This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

Exemplar-based Recognition of Human-Object Interactions

Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, Shaogang Gong, and Tao Xiang

***Abstract*—Human action can be recognised from a single still image by modelling human-object interactions (HOI), which infers the mutual spatial structure information between human and the manipulated 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 probabilistic density functions describing spatially how a person is interacting with a manipulated object for different activities. Specifically, a new framework consisting of an exemplar-based HOI descriptor and an associated matching model is formulated for robust human action recognition in still images. In addition, the framework is extended to perform HOI recognition in videos, where the proposed exemplar representation is used for implicit frame selection to negate irrelevant or noisy frames by temporal structured HOI modelling. Extensive experiments are carried out on two image action datasets and two video action datasets. The results demonstrate the effectiveness of our proposed methods and show that our approach is able to achieve state-of-the-art performance, compared with several recently proposed competitors.**

***Index Terms*—human-object interactions, action recognition, exemplar modelling**

I. Introduction

RECENTLY the problem of recognising action of a person who is manipulating objects from a single image or video has received increasing interest [1], [2], [3], [4]. In this context, the action is regarded as the Human-Object Interaction (HOI). For example, the action "playing a guitar" can be described as a human holding a guitar under some certain poses. Therefore, interactions are the main characteristic of an action as well as the actor's pose and manipulated object's appearance. 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 as a mutual context model that joins object detection and human pose estimation (i.e. the posture information) together [5]. In contrast, the spatial relationship concerns more about the relative geometric information, e.g. the relative position and overlap between human and object that join human detection or annotation and object detection together [6], [2], [7], [8]. In addition, global context that describes holistic semantic information about HOI has been exploited recently to assist HOI modelling [4].

However, most of existing HOI modelling approaches rely heavily on explicit human pose estimation [5] for co-occurrence modelling or directly use locations of human and object to explore the spatial information in HOI representation [6], [2], [7]. Nevertheless, the problem of detecting objects,

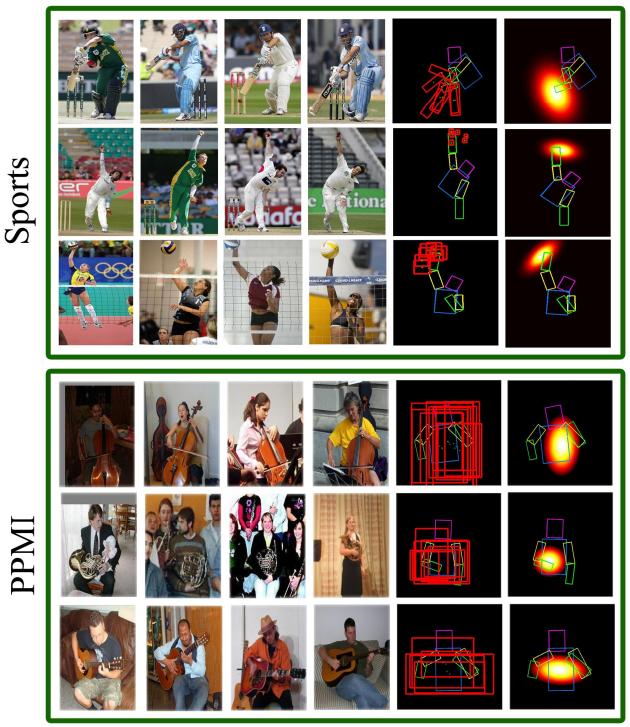


Fig. 1. Graphical illustration of spatial pose-object interaction exemplar computation in the Sports and PPMI datasets [2]. Each row shows example of an exemplar. Columns 1-4 show four images represented by the same atomic pose. Column 5 shows the manipulated object locations overlapped with corresponding atomic pose. Red boxes indicate objects. Column 6 shows the exemplars. Hotter colors indicate larger response.

This research was supported by the National Natural Science of Foundation of China (Nos. 61102111, 61173084, 61472456, U1135001), the 12th Five-year Plan China S&T Supporting Programme (No. 2012BAK16B06), Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, and in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265. (Corresponding author: Wei-Shi Zheng)

J. Hu is with the School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, China, and SYSU-CMU Shunde International Joint Research Institute, Shunde, China. E-mail: hujianf@mail2.sysu.edu.cn

W.-S. Zheng and J. Lai are with the School of Information Science and Technology, Sun Yat-Sen University, Guangzhou, China. W.-S. Zheng is also with Guangdong Provincial Key Laboratory of Computational Science, China. J. Lai is also with Guangdong Province Key Laboratory of Information Security, China. E-mail:wszheng@ieee.org and stsljh@mail.sysu.edu.cn

S. Gong and T. Xiang are with the School of Electronic Engineering and Computer Science, Queen Mary University of London.E-mail:s.gong@qmul.ac.uk and t.xiang@qmul.ac.uk

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

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 as a challenging problem. Therefore, the performance of existing approaches can be hindered by inaccurate human/object detection and pose estimation.

In this paper, we aim to 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 different instances of a human action class, the manipulated object, 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 Figure 1). Therefore, the configuration of pose and object can be viewed as an exemplar for describing the action in terms of the interaction between human and object. This type of exemplars is termed as *spatial pose-object interaction exemplar* in this work.

The introduced spatial pose-object interaction exemplar is represented as a density function that describes how likely an object appears at a certain location in an image with respect to a certain (atomic) pose of a position in the image. Some examples of spatial pose-object interaction exemplars can be found in Figure 1 column 4 and Figure 2. By measuring the response of different exemplars in that image, we obtain a probabilistic model for the mutual structure information between human and object in our exemplars. It can greatly alleviate the effect of inaccurate human and object detection and avoid explicit estimation of human pose. Since the spatial pose-object exemplar only captures the geometric information between pose and object, for a more comprehensive representation of HOI, we also consider the appearance information of pose and object, which is complement to the spatial information. Hence a HOI descriptor is formulated by combining the response of spatial pose-object exemplar, response of appearance models of pose and object, and global context. Furthermore, in order to quantify the importance of the HOI descriptor and thus select different components in the proposed HOI descriptor, a new action-specific ranking-based matching model is formulated.

Though being originally designed for HOI recognition from still image, the proposed exemplar-based HOI representation is extended from still image to video by exploring temporal structured HOI modelling. We show that by applying the proposed exemplar-based HOI on video frames, frame selection can be conducted implicitly to alleviate the effect of irrelevant or noisy frames during HOI modelling, because the exemplar model response can indicate how likely any target HOI would appear in the frame.

We evaluate the effectiveness of our approach on four benchmark HOI datasets, including two still image datasets and two video datasets. Our results on the two image sets (sports [9] and PPMI dataset [4]) demonstrate that our approach is able to produce state-of-the-art performance, compared with the most recently proposed competitors on still images. Moreover, we also test our method on two video datasets: Gupta video dataset [10] and SYSU activity dataset; the latter is a new dataset introduced in this work. The results validate the feasibility of the extension of the proposed HOI modelling from still image to video and show that it outperforms a recent state-of-the-art method [11].

Our contributions are as follows: 1) formulating a novel HOI representation based on a set of pose-object interaction exemplars; 2) quantifying and selecting the most discriminant components of the proposed HOI descriptor for action recognition by an action-specific matching method; 3) exploring a unified framework for HOI action classification for both still image and video; 4) making available a challenging new HOI video dataset, namely the SYSU activity dataset.

II. Related work

Human action/activity recognition from video is a broad and active area of research and has been widely studied in the last decade. It is out of the scope of this paper to give a complete review on all the works in this area. Readers are recommended to refer to [12], [13] for the latest and most comprehensive reviews. Inspired by the fact that humans can recognise action from a single still image, recently computer vision researches have attempted to model actions from static images. The actions studied are typically characterised by specific human poses or their combinations with scene context [14], [15], [6], [16], [17]. The human pose cue is often captured by a pictorial structural model [18] or 'poselet' [19]. To avoid unreliable pose estimation, algorithms in [20], [21], [22], [23], [24] treat action recognition as a typical image classification problem without explicitly modelling the pose of human of interest. There are also some works that combine the action recognition in still image and video together [25], [26], [27].

In this work we are interested in one specific type of actions: human-object interactions (HOI) [5], [28], [29], [30], [31]. These actions are often defined by the interactions between a human and an object [2], [5], [9]. Most existing HOI modelling approaches focus on modelling the co-occurrence or spatial relationship between human and the manipulated object. In [9], various perceptual tasks, such as action recognition and object recognition, are integrated to understand human-object interactions. A weakly supervised learning method is developed to model the spatial information between human and object in [2]; [6] proposes a representation for HOI classification using poselet [19]; [7], [14] train a set of visual phases (complex visual composites such as "a person riding a horse") detectors for HOI recognition and detection. These methods exploit the spatial relationship between the locations of person (bounding box either detected or manually annotated) and object (detected) to model HOI. Different from encoding the relationship between human and object, [4] identifies human and object interactions using a global image feature called "Grouplet" to capture the structured information of HOI. In [32], complex interactions are modelled by using velocity histories of tracked keypoints. Recently, some works are proposed to model HOI in RGB-D videos [33], [34], [35]. For instance, [29] presents a method for categorising manipulated objects and tracking 3D articulated hand pose in the context of each other in order to recognise the interactions between human and object. In addition to explicitly modelling the spatial relationship

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

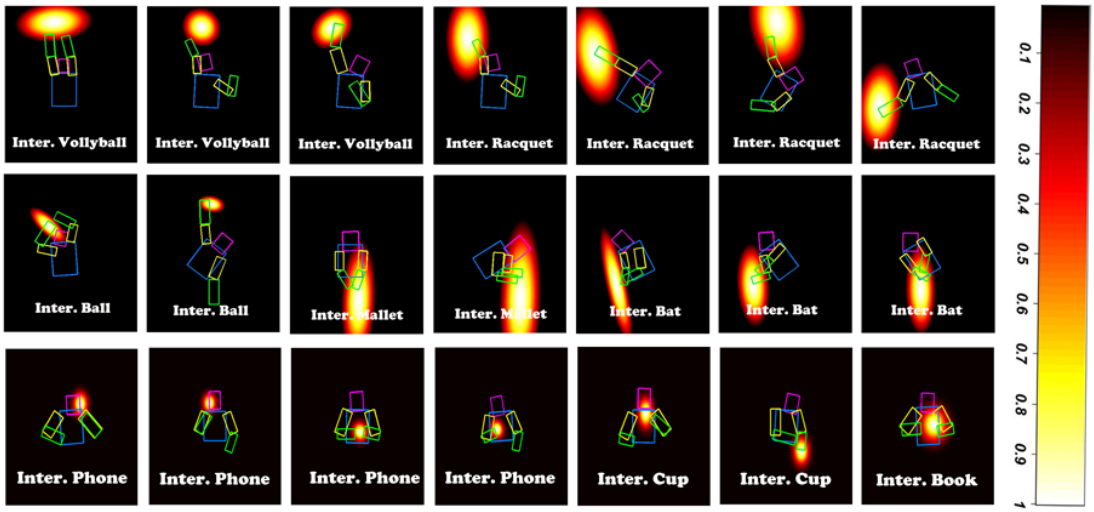


Fig. 2. Some examples of spatial pose-object interaction exemplar. All annotated boxes in an image constitute an atomic pose, where different parts are discovered and marked with different colours. The bright areas visualise the spatial distribution of the manipulated object.

between human and object, the relative motion of object w.r.t human is also exploited to describe their interactions in [11].

One of the key limitations of most existing HOI modelling approaches is that they all heavily rely on accurate and robust object detection and pose estimation. Both are far from being solved therefore hindering the performance of exsiting approaches. In order to overome this limitation, an exemplar-based approach is proposed in this paper to model the interactions between a person and manipulated object. Although exemplar-based modelling has been applied to a variety of visual recognition problems including scene recognition [36], object detection [37], and pose estimation [38], few works exploit exemplar-based modelling for HOI recongition. Moreover, the use of exemplar in existing work is focused on transferring useful information extracted from meta-data to a new set of data. This is very different from our objective, that is, to develop an exemplar-based representation for HOI. Recently, exemplar-based methods have been proposed for action recognition on still images [15] and video [39]. Again, the purpose of using exemplars in [15], [39] is very different from ours – examplars are used for selecting a set of representative samples for each class, rather than for representing HOI by modelling the relative spatial relationship between a human and an object probabilistically; in addition and crucially, our approach does not rely on local feature point estimation, explicit pose estimation or depth information estimation. Recently, Yao et al. [40] proposed to model object functionality that describes each type of human-object interactions. This work is related to our exemplar-based approach in that it also explores the possible interactions between human pose and object and treats them as prototypes. However, Yao's modelling has a very different objective – to refine 3D pose estimation and object detection in an iterative way, rather than assisting HOI recognition. In addition, different from Yao's work, our work does not need to explicitly estimate human pose or to perform any 3D modelling. Moreover our approach is not iterative, therefore can be less expensive and more readily extendible to dealing with exemplar HOI modelling in videos.

A preliminary version of parts of this work was presented in [41]. However, in this work, we further extend our preliminary model for identifying HOI in video by developing a novel temporal structured HOI descriptor and introducing a new challenging video dataset (SYSU activity) for studying the interactions between human and manipulated object. In addition, we provide more analysis and experiments about the matching model in our approach. More experimental results are reported to demonstrate the effectiveness of our modelling.

III. HOI Exemplar for Still Image

Given an image of HOI action, our goal is to both identify and describe the interactions between human of interest and his/her manipulated object involved in the image. We achieve this by proposing an exemplar-based approach, which first automatically discovers a set of human-object spatial interaction exemplars to represent the interactions, and second employs a matching model to combine different cues derived from human, object, scene context and the interaction between human and object. Specifically, our approach consists of two main parts: 1) a new exemplar-based HOI descriptor (Sec.III-A∼Sec.III-D); and 2) a matching model for learning the optimal combination weights of all cues in the proposed HOI descriptor (Sec. III-E).

A. Learning Atomic Poses

Instead of explicit human pose estimation, our modelling is based on the use of a set of atomic poses [5] learned from training data. Atomic poses are representative poses that often occur in a specific type of HOI action. Given the learned atomic poses, each pose involved in the action is associated to the closest (visually 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 the method in [5]. In particular, during training the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

location of each human body part is annotated. To compute 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 the Affinity Propagation (AP) clustering method [42], [43]. The computed cluster centres $H = \{H_1, H_2, ..., H_N\}$ form the dictionary of atomic poses, that is, each cluster represents an atomic pose. Some examples of the learned atomic poses from a sports dataset are illustrated in Figure 1 and Figure 2. The advantage of using the AP method is that we do not need prior knowledge on the number of atomic poses N – it is determined automatically.

B. 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 object. Our idea of exploring interaction exemplar is inspired by the observation that the locations of the manipulated object are constrained by human's location, pose and type of action. For example, if a man is playing volleyball as illustrated in the first picture of Figure 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. To this end, we formulate a distribution function $G(\mathbf{x})$ to describe the likelihood that a manipulated object would appear at location $\mathbf{x}$ around a human body for a specific spatial pose-object interaction. In this work, we call such a distribution *Exemplar*. By utilising the distribution modelling, we are able to describe the interaction between pose and object probabilistically, rather than directly using the label information or precise coordinates of object and human body as features for inference.

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

Dictionary Construction. We assume the distribution of each 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 body in an action. That is, we formulate the density function for an exemplar as

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

For each training sample $Q_{nk} \in \mathcal{Q}$, we denote its corresponding atomic pose as H_n and its manipulated object as O_k. We aim to learn a spatial pose-object interaction exemplar $G(\mathbf{x})$ which describes how likely O_k should be located 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. Here, 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 being associated to a pair of 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 the i^{th} sample in $\mathcal{Q}_{nk}$. Then the estimation of $(\boldsymbol{\mu}_{nk}, \Sigma_{nk})$ is given by

$$\boldsymbol{\mu}_{nk} = N_{nk}^{-1}\sum_{i=1}^{N_{nk}}\tilde{\mathbf{L}}_i, \Sigma_{nk} = N_{nk}^{-1}\sum_{i=1}^{N_{nk}}(\tilde{\mathbf{L}}_i - \boldsymbol{\mu}_{nk})(\tilde{\mathbf{L}}_i - \boldsymbol{\mu}_{nk})^T \quad (2)$$

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 atomic 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 measures the 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 Figure 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 exemplar to describe the interaction between them. In addition, from this figure, we can observe that our spatial pose-object interaction exemplar can capture some semantic information about how the actor is manipulating the object.

C. Inferring Spatial Pose-Object Interaction Using Exemplars

After constructing the exemplar dictionary, the learned dictionary is employed to encode the interactions involved in a probe image. Without an explicit pose estimation, given a test/probe image, we first search for the most similar atomic poses. Based on the nominated atomic poses, the model selects the candidate exemplar in the dictionary and computes the response of probe HOI against the exemplars. Finally the model forms a code vector for each probe HOI consisting of all the response of the exemplars in the dictionary. In the following, we detail the whole process which is also illustrated in Figure 3.

*1) **Nominating Similar Atomic Poses**:* For each probe HOI image, we nominate the most similar atomic poses defined in the spatial exemplar dictionary. For each detected person P in the probe HOI image, we first score each training image with $Sim(P, P^i)$, where $Sim(P, P^i)$ is a function that measures

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

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 in the experiments (Sec. V). To compute $Sim(P, P^i)$, we compute the inverse of the distance between their feature representations encoded by pyramid histogram of words (PHOW) [44]. For obtaining the PHOW feature, we extract dense SIFT features, learn a vocabulary of size 512, and finally compute the histogram under three pyramid levels. Here, we further expand each PHOW feature to a vector of dimension 32,256 using an approximated kernel map for the Chi-Square kernel [45]. As stated in [2], pyramid image features can capture soft pose information. Note that only upper body is considered for learning the atomic poses, since sometimes only upper body is visible for the person of interest in an HOI action.

*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 pre-trained 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 $(x_t, y_t$ are the coordinates of human centre, and w_t and h_t are the width and height of the torso respectively.

Second, for the k^{th} object type O_k, we detect objects of this type in the image and predict the most likely existing location (x, y), 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. The aligned exemplar $\tilde{G}_{nk}(x, y)$ gives 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 Figure 4 column 2 for examples of $\tilde{G}$). Then the exemplar response can be defined as

$$\mathbf{I}(n, k) = \tilde{G}_{nk}(x_o, y_o). \quad (4)$$

We compute Eq. (4) for each selected candidate atomic poses and each object type. We set the entries corresponding to non-selected atomic pose to zero. Finally, the obtained matrix is then reshaped as a vector $\mathbf{I}$ (with a slight abuse of notation).

D. Exemplar-based HOI Descriptor

The exemplar response vector $\mathbf{I}$ as described in the last section only captures the mutual spatial structure information, i.e. the probabilistic geometric information between human and object. It does not capture appearance information about the pose and the object, which is also important for describing HOI, because different types of pose-object exemplar may have similar spatial response. Hence, further including the pose appearance feature $\mathbf{P}$ and object detection vector $\mathbf{O}$ can reduce the ambiguity of exemplar modelling. So, the combination of $\mathbf{I}$, $\mathbf{P}$, and $\mathbf{O}$ provide complementary information to each other and form a main part of our exemplar descriptor, where $\mathbf{I}$

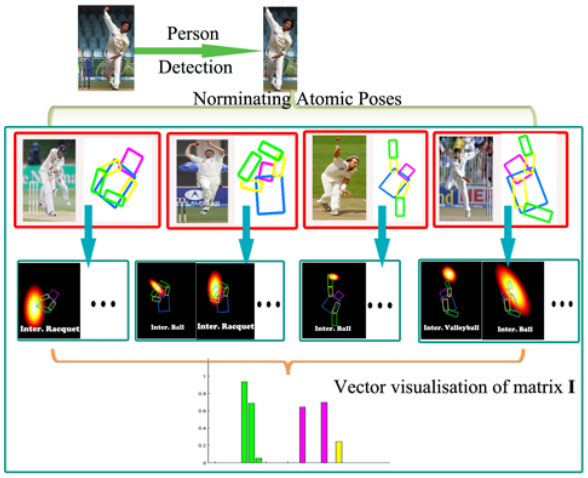


Fig. 3. A graphical illustration of how exemplar response is computed in still images. The last row is a vector visualisation of matrix $\mathbf{I}$ in Eq. (4), where for better visualization, bars associated to different manipulated objects are marked with different colors: cricket bat (red), cricket ball (green), croquet mallet (blue), tennis racket (magenta), and volleyball (yellow). From the final representation, we can see that the actor is manipulating a tennis racket or cricket ball.

indicates the spatial interaction response and $\mathbf{P}$ and $\mathbf{O}$ forms the appearance interaction response. In addition, similar to existing approaches as [5], [2], [9], [3], 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 $\mathbf{C}$ around a person, i.e.

$$\mathbf{H} = [\mathbf{I}; \mathbf{P}; \mathbf{O}; \mathbf{C}] \quad (5)$$

Compared to existing HOI descriptors [5], [2], [3], [9], [11], our model differs primarily in the use of the spatial exemplar response $\mathbf{I}$ and the relaxation of the assumption about accurate object and pose estimation. Note that not all parts of the descriptor are equally informative for representing HOI. In the next section, a matching model is proposed in order to implicitly perform feature selection.

Some examples of our HOI descriptor are shown in Figure 4. It can be seen that our descriptor captures information about what (manipulated object), who (person of interest) and how (in which way a person manipulates the object) to provide a comprehensive representation of HOI action in an image. More specifically, given an input image, we first predict its activity class and the manipulated object type. We then locate the person (who) and manipulated object (what) of interest. Some visualisation can be found in Figure 4 columns (c) and (g). The yellow and magenta dashed rectangles indicate object and person, respectively. The object type is provided beside object box. Results of the normalised exemplar are shown in Figure 4 columns (b) and (e). We can observe that the normalised exemplar can provide a strong prior of the position of manipulated object. Figure 4 columns (d) and (g) show the interaction component (i.e $\mathbf{I}$ in Eq. (5)) of our HOI descriptor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

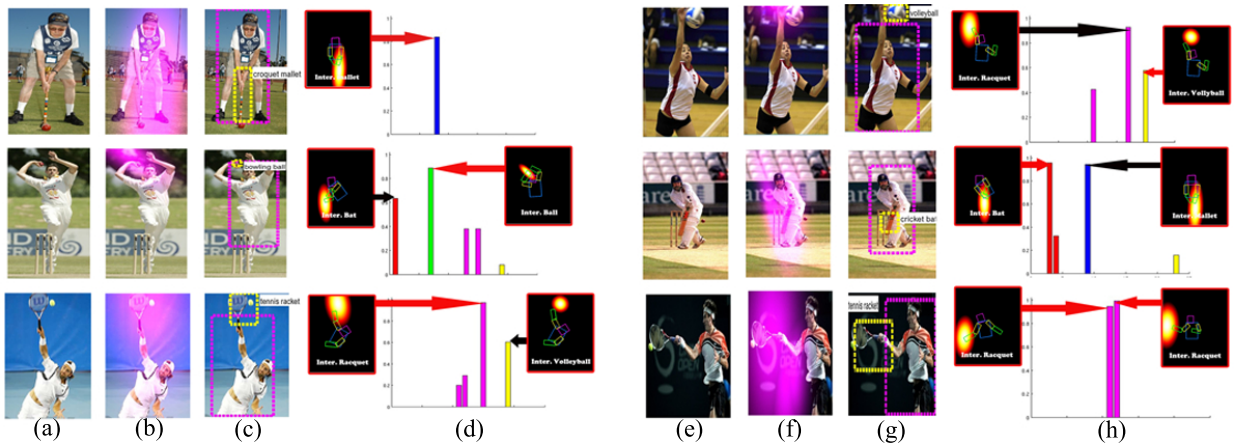


Fig. 4. Examples our HOI descriptor. Examples from two action classes are shown. In particular, we visualise the pose-object exemplar response vector (**I** in Eq. (5)) in columns (d) and (h). Exemplars with large response value (>0.5) are presented besides the bar graph. Bars that represent different objects are plotted 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 the predicted action type. These examples demonstrate that our exemplar response captures semantic information about the HOI action regarding how a person manipulates an object of what type.

for the given image.

We also note that some incorrect exemplars may have a large response value. Hence, the interpretation can be better if the decisions on the manipulated object type and activity class obtained by our action classifier introduced later are used to filter out the incorrect templates. That is why scene context and pose information can facilitate our activity interpretation as well as recognition. For instance, as shown in the right of Figure 4 row 2, there are two exemplars with large response (> 0.5) values. One (left) indicates that the actor is manipulating a cricket bat, while the other (right) shows that he is interacting with a croquet mallet. It is intuitive to remove the right exemplar, because its object type is not consistent with the manipulated object type. Finally the left exemplar is selected for interpreting the activity.

E. Matching Model

We have four components for each HOI descriptor, and each component can provide different cues for HOI action recognition. We wish to quantify all four cues in our HOI descriptor for our activity analysis so that the final matching is based on a weighted combination of the four. Intuitively, given sample Q_i, we formulate the assignment according to the following score function which is a weighted combination of probability score of each component:

$$\begin{aligned} f(a_z|Q_i) =& \alpha_z p(a_z|\phi_P(Q_i)) + \beta_z p(a_z|\phi_O(Q_i)) \\ &+ \gamma_z p(a_z|\phi_C(Q_i)) + \zeta_z p(a_z|\phi_I(Q_i)) \end{aligned}$$

$p(a_z|\phi_P(Q_i)), p(a_z|\phi_O(Q_i)), p(a_z|\phi_C(Q_i)), p(a_z|\phi_I(Q_i))$ indicate the probability of assigning label a_z to a given sample Q_i from the perspective of human pose, manipulated object type, global context and spatial interaction cues, respectively. Entry $\phi_*(Q_i)$ is the corresponding descriptor extracted from sample Q_i as discussed in the above section. Parameters $\alpha_z, \beta_z, \gamma_z, \zeta_z$ are employed to encode the reliability of the corresponding assignments (i.e. the weights). In order to obtain $p(a_z|\phi_*(Q_i))$, we feed all $\phi_*(Q_i)$ into an one-vs-all discriminative classifier and use the learned classifier to predict the probability.

For inference, we assign a label to sample Q_i by

$$a^* = \arg\max_{a \in \mathcal{C}} f(a|Q_i) \tag{6}$$

Intuitively, we can obtain a naive matching model by simply averaging over all the four components i.e. setting all the parameters $\alpha_z, \beta_z, \gamma_z, \zeta_z$ to be 1. However, in order to get the optimal prediction, we wish that the correct prediction has higher score than the incorrect ones. Hence we learn the parameters by solving the following optimisation problem

$$\begin{aligned} &\min \frac{1}{2}\sum_{z=1}^{Z}(\alpha_z^2 + \beta_z^2 + \gamma_z^2 + \zeta_z^2) + \frac{1}{vM}\sum_{i=1}^{M}\xi^i \\ &s.t. f(a_i|Q_i) \geq f(a|Q_i) + 1 - \xi_i, \forall i = 1, 2, ...M, a \in \mathcal{C} \ /\{a_i\}, \end{aligned}$$

where a_i is the ground truth label of the i^{th} training sample, M and Z represent the training sample number and action class number, respectively. $\sum_{z=1}^{Z}({\alpha_z}^2 + {\beta_z}^2 + {\gamma_z}^2 + {\zeta_z}^2)$ is a regularization term to avoid over-fitting. By denoting $(\alpha_z, \beta_z, \gamma_z, \zeta_z)^T$ and $(p(a_z|\phi_P(Q_i)), p(a_z|\phi_O(Q_i)), p(a_z|\phi_C(Q_i)), p(a_z|\phi_I(Q_i)))^T$ as $\mathbf{w}_{a_z}$ and $\mathbf{s}_{a_z}$, respectively, we can get a large-margin optimisation problem as

$$\begin{aligned} &\min \frac{1}{2}\sum_{z=1}^{Z}\|\mathbf{w}_{a_z}\|^2 + \frac{1}{vM}\sum_{i=1}^{M}\xi^i, \\ &s.t. {\mathbf{w}_{a_i}}^T \mathbf{s}_{a_i}^i \geq {\mathbf{w}_a}^T \mathbf{s}_a^i + 1 - \xi^i, \ \xi^i \geq 0 \\ &\qquad \forall i = 1, 2, ...M, a \in \mathcal{C} \ /\{a_i\}, \end{aligned} \tag{7}$$

where s_*^i indicates the confidences on how likely the sample is from class $*$ and v is a parameter to control the trade-off between training error minimisation and margin maximisation. The larger the v is, the more points are allowed to lie inside the margin. We set v to be 0.07 in this work.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

Solving the above quadratic programming problem directly is not straightforward. However, inspired by [46], we next show that any one-class SVM solver can be used for computation by applying a simple transformation in Prob. (7). 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 $\mathbf{S}^i_{aa_i} = \phi(a_i) - \phi(a)$, where $\mathbf{0}$ is a zero vector. Then the optimisation problem (7) can be rewritten as

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{vZ}\sum_{i=1}^{M}\xi^i, \\ &s.t. \mathbf{w}^T\mathbf{S}^i_{aa_i} \geq 1 - \xi^i, \xi^i \geq 0 \\ &\qquad \forall i = 1, 2, ...M, a \in \mathcal{C}\ /\{a_i\}. \end{aligned} \tag{8}$$

Let $\mathbf{S} = \{\mathbf{S}^i_{aa_i}\}_{i=1,2,...M, a\in\mathcal{C}\ /\{a_i\}}$. Note that the solution of Prob.(8) would linearly separate $\mathbf{S}$ from the origin with maximum margin, which means that we can solve the proposed optimisation problem using any general SVM solver that is suitable for one-class SVM, although Prob.(7) is not a regression problem.

The formulation of our matching model (Prob.(7)) is related to RankSVM. However, the relative comparison is essentially different. In the constraint given by Prob.(7), its weighting vector is defined differently on each side of the inequality, unlike in the case of RankSVM. As a result, this matching model measures direct comparative scores from different weighting models, whilst the RankSVM model computes relative comparison between intra- and inter-class scores. In this way, a specific rather a common weighting vector is learned for each HOI action. Moreover, our feature mapping is defined by the outputs of a set of weak classifiers rather than the original image features. Thus, it can dramatically reduce the memory usage for learning the matching model. In addition, it is also possible to fuse different types of information using a multiple kernel learning (MKL) framework. However, our matching model aims to directly optimise the weighting in a discriminative framework whilst models such as the generalised MKL [47] indirectly do so. Consequently our model has the potential to learn a better weighting of different cues for our recognition task. In our experiments, we compare the proposed matching model with RankSVM, a generalised multiple kernel learning (MKL) approach, and other related methods to show that the proposed matching model is superior to the alternatives.

IV. HOI Exemplar for Video

In this section, we show that our exemplar modelling can be extended to recognise human-object interactions in video. We would extend the work by introducing a temporal structured HOI representation to encode the temporal ordering between exemplar spatial interaction responses derived from individual video frames.

Here, we treat video clip as an ensemble of still images (frames) $\{v^f\}_{f=1,2...F}$. We can calculate an exemplar response vector for each frame using the method described in the previous section. We denote the exemplar response vector of the f^{th} frame as $\mathbf{I}^f$. Our goal is thus to aggregate them into a single vector of fixed dimensionality as the video's final spatial interaction representation. To consider the temporally structure of video data, a $L-level$ temporal pyramid is constructed by repeatedly partitioning the video clip into increasingly finer sub-segments along the temporal dimension, in a way similar in spirit to spatial pyramid [48] used in image representation. For the level $l(l = 0, 1, ..., L-1)$, 2^l sub-segments are constructed. Interaction response vectors of frames found in each sub-segment are pooled together. This results in a total of 2^L sub-segments and the corresponding representations. Then the concatenation of all the 2^L representations forms our final video semantic spatial interaction response $\mathbf{I}$. This process is illustrated in Figure 5.

Now we describe in details how to pool the vectors found in each sub-segment together. It is desirable that the pooling method is robust to noise and independent of video length F. Intuitively, a MAX or SUM pooling method can be selected to pool them together, whose effectiveness has been demonstrated in many tasks including image classification [49] and scene classification [50]. However, the previous approaches typically ignore one important fact, that is, a lot of irrelevant video frames do not correspond to any type of action, since MAX and SUM would perform the pooling by utilising all frames even when some of them are irrelevant or noisy. To solve this problem, we introduce a weighted pooling method to pool all the response together. The weights are set simply to the sum of the response. Our pooling method can thus be formulated as

$$\mathbf{I}_{lj} = \frac{\sum_{f\in H_{lj}} s^f \times \mathbf{I}^f}{\sum_{f\in H_{lj}} s^f} \tag{9}$$

where s^f is the sum of the response entries in $\mathbf{I}^f$, H_{lj} is the index set of frames found in the j^{th} subsegment of level l, $j = 1, ..., 2^l, l = 0, 1, ..., L-1$. It is worthy noting that for the redundant frames, the responses are usually small, so our exemplar response of each frame can indicate how likely the frame consists of any HOI of interest and the proposed pooling method can therefore serves as an implicit frame selection method over all the video frames by filtering out the noisy or irrelevant frames. As evident in the third row of weight plots in Figure 5, the third, fourth and fifth illustrated frames with large weights (brighter colour) are more relevant to the action of 'answering a phone', whilst the first, second and sixth frames with smaller weights (darker colour) are noisy and less informative about the action. We further show in Table VI that our proposed weighted pooling performs better than the traditional ones, so as to alleviate some inaccurate exemplar response (e.g. exemplar related to the spray has a sharp response in the fifth illustrated frame, while it is not in the others).

Finally, the final representation $\mathbf{I}$ is fed into our matching model in Sec. III-E to obtain the final prediction along with the other components (person descriptor, object detection score and global contextual information) of our full HOI descriptor.

V. Experiments

Two types of experiments are carried out, i.e. action recognition in still image and video respectively. For recognition in still image, we evaluate our method on two benchmark

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

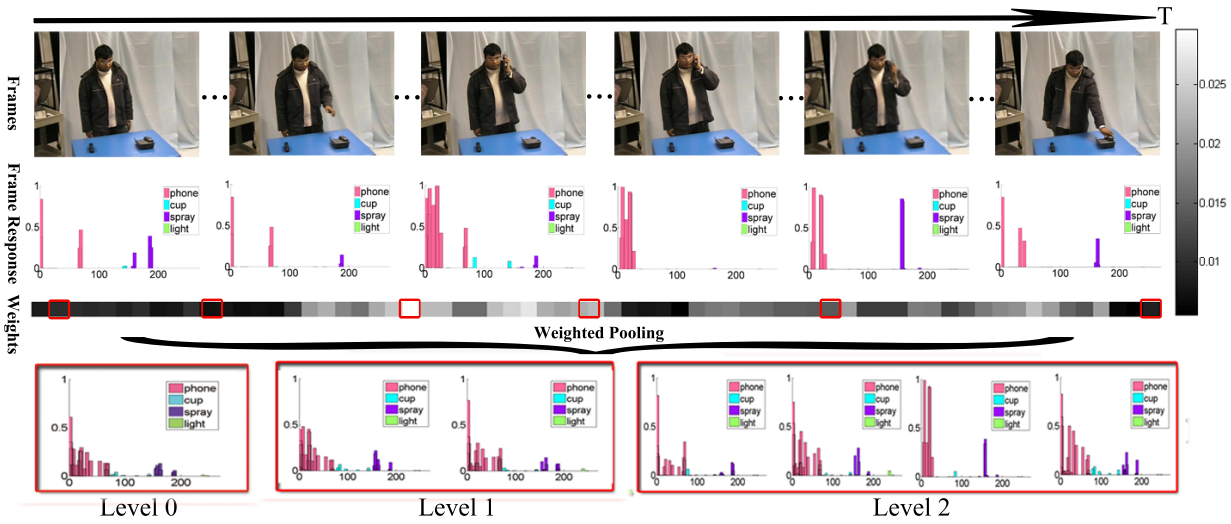


Fig. 5. Graphic illustration of how exemplar response is computed in video. The first row presents six video frames in a video clip; the second row presents the spatial exemplar HOI response in the corresponding frame; the third row illustrates the weights of all the frames in the clip, a darker colour means a small weight and a whiter one means a large one, where the ones with a red box corresponds the illustrated frames in the first row. The fourth row illustrates the spatial interaction response $\mathbf{I}$ in a pyramid modelling. For better visualisation of the response vector, entries corresponding to different manipulated objects are shown in different colours.

Method	Yao [5]	Desai [3]	Prest [2]	Gupta [9]	Our Model
Accuracy (%)	87	82.5	83	78.9	**92.5**

TABLE I
COMPARISON ON THE SPORTS DATASET.

datasets: a sports data set [9] and a people-playing-musical-instrument (PPMI) data set [4]; for recognition in video, we test our proposal on two video datasets: Gupta video dataset [10] and a newly collected video dataset named Sun Yat-sen university HOI action(SYSU action) dataset.

A. Recognition in Still Images

1) Experimental Settings: The descriptor, detection and parameter setting are described below.

HOI descriptor details. For global contextual information $\mathbf{C}$ and person descriptor $\mathbf{P}$, we extracted spatial pyramid histogram of words (PHOW) from the person of interest and whole image respectively. For object detection vector $\mathbf{O}$, we ran each object detector and concatenated the top detection score of each object type.

Detection. Although our model does not rely on accurate object detection and pose estimation, detection is still needed. We detected human, body parts and objects using the deformable part model [51] with two mixture components for the sports dataset and PPMI dataset. To train detectors of human, head, torso and upper body, the ground truth bounding boxes in the sports dataset and PPMI were used to generate positive examples, and the negative samples, namely the background images were generated from PASCAL VOC12 [52]. We followed [51], [2] to obtain detection of person across a variety of poses. Regarding the detection of other objects, for each object type, we used the corresponding trained detector to obtain the centre location (x_o, y_o) of the object. Note that in order to rely less on the object detection accuracy, we only used the detected location to represent object without using its scale at this step.

Parameter Settings. The number of candidate exemplars for computing the exemplar spatial interaction response in Sec. III-C2, namely the parameter S was set to 3 for the sports dataset and 20 for PPMI, which are about one fourth of the number of the learned atomic poses. Its effects will be evaluated in Sec. V-C1.

2) Results on the Sports Dataset: The sports dataset consists of 300 images of six HOI activities (tennis-forehand, tennis-serve, volleyball-smash, cricket-bowling, cricket-defensive shot, croquet-shot). Figure 1 shows some interaction examples. As shown, the involved poses and human-object interactions in this set exhibit small intra-class variation. We followed the same experiment setting as in [5], [2], [3], [9], where for each activity 30 images were selected for training and 20 were selected for testing. As in [1], only five object classes, namely cricket bat, bowling ball, croquet mallet, tennis racket, volleyball, were employed to model and evaluate HOI for action recognition.

We compare our method with the following competitors: ***Yao*** [5], ***Prest*** [2], ***Desai*** [3] and ***Gupta*** [9]. All these methods utilise pose, object, relation between pose and object and contextual information. In contrast to our method, they either directly use the estimated locations of people and human as features [2], [9] or require explicit human pose estimation [5], [3].

Table I shows the results. It can be seen that the proposed model achieves the best performance and outperforms the state-of-the-art [5] by 5.5%. It also improves over [9], [3] and [2] by 13.6%, 10% and 9.5%, respectively. The confusion matrix of our model is shown in Figure 7(a). We can observe that our model achieves perfect results on the actions of cricket-batting and croquet. It is noted that serious false detection and occlusion can still affect the performance of our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

Method	SPM [48], [4]	Grouplet [4], [5]	Yao [5]	Our Model
Accuracy (%)	41.8	-	-	**49.34**
mAP (%)	39.1	42	48	**47.56**

TABLE II
COMPARISON ON THE PPMI DATASET. '-' INDICATES THAT THE CORRESPONDING RESULT HAS NOT BEEN REPORTED BEFORE.

model. For example, for classification of volleyball-smash and cricket-bowling in the sports dataset, our model achieves lower classification accuracy ($\leq$90%) (see Figure 7(a)). In particular, 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 there are some audiences who do not play volleyball existing in the background. Nevertheless, compared to the other methods, it is relatively less sensitive to the detection errors, resulting in superior performance.

3) Results on the PPMI Dataset: For PPMI, there are twelve musical instruments, and each image contains people playing/holding an instrument. The dataset contains 2400 images for training and 2400 images for testing [4]. As can be seen from Figure 1, there is a much greater degree of intra-class variability in this dataset as compared to the Sports dataset. In addition, different classes are visually much more similar in this dataset (e.g. a number of musical instruments are played with a similar body pose). We follow the setting in [5] to select a subset which gives us 2175 images for training and 2035 images for testing.

In this experiment, we evaluate different methods on the 24-class classification task. Since the annotations of the dataset used in [5] are not available, we have taken the best efforts to re-annotate this dataset in the same way as what was done for the sports dataset. Specifically, for each training image, we annotated manipulated object and six body parts, including head, torso, left upper arm, left lower arm, right upper arm and right lower arm.

The comparative results are presented in Table II[1]. From this table, it can be seen that our model achieves 49.34% in average accuracy and 47.56% in mAP (mean average precision), with an improvement of 6% to 8% over SPM [48] and Grouplet [4]. The proposed model yields very similar result compared with the state-of-the-art Yao's method [5] on this dataset in terms of mAP.

B. Recognition in Video

We further test our method on two video datasets: *Gupta Dataset* [9] and *Sun Yat-sen university (SYSU) action Dataset*, where the latter is a newly collected one.

1) Experimental settings: The descriptors, detection and parameter setting are as follows.

HOI descriptor details. For the person descriptor $\mathbf{P}$, we extracted holistic features from the clouds of interest points around a person accumulated over multiple temporal scales using the method presented in [53], where all parameters were set by following the author's suggestion. For global contextual information $\mathbf{C}$, we computed the dense optical flow features and quantised them into a fixed number of discrete 'bins' according to their coordinates. In our experiment, 60×60 bins are employed. This results in a histogram vector of 3600 dimensions. For the object detection vector $\mathbf{O}$, we ran each object detector on all video frames and concatenated the top 10 detection scores of each object type.

Detection and Settings. We need to track the actor and the manipulated objects. Similar to other related work [11], [54], we tracked the locations of object and person using a series of pre-trained detectors. Our detectors were trained on the training set and VOC2012 (used for generating negative examples) using DPM [51] with two mixture components. The selected candidate exemplar number was set to 12 and 20 for Gupta and SYSU datasets, respectively. This free parameter will be further evaluated in Sec. V-C1. Other parameters were set to the same as those in still images.

Video annotation. For each video clip in the training set, we annotated each frame with seven bounding boxes indicating locations of head, torso, upper left arm, down left arm, upper right arm, down right arm and the manipulated object. Rather than manually labelling each frame, we took the temporal continuity in video data into consideration and utilised the annotations of the neighboring frames to facilitate the current frame's annotation. More specifically, in the training stage, we annotated the video every five frames and then adapted the manual annotations to its neighbouring un-annotated frames. For adaptation, given a frame F_c to be annotated and the annotation l_p (i.e. coordinates) of its neighbouring frame F_p, the candidate annotation is obtained as follows

$$(\sigma^\star, \mathbf{s}^\star) = \arg \min_{\sigma \in \{3,5,10\}, \mathbf{s} \in \triangle} \|G(F_c, l_p + \sigma \mathbf{s}) - G(F_p, l_p)\|^2 \tag{10}$$

where $G(*)$ indicates the appearance features extracted from the corresponding frame and bounding box, $\triangle = \{(0,0); (1,0); (1,1); (0,1); (-1,1); (-1,0); (-1,-1); (0,-1); (1,-1)\}$. Therefore, for each un-annotated frame, we can get two candidate annotations according to its two neighbouring frames. Finally, the final annotation is given by their weighted average location whose weights are obtained by the minimal distance in Eq. (10). Some annotation examples can be found in Figure 6.

Compared Methods. We compare our method with three methods: *IP* [53], *3DHOG* [55] and *Explicit Model* [11]. As a widely used baseline, *IP* employs a spatio-temporal interest point based method to predict the action type of a given video. In the implementation of *IP*, we followed the parameter setting in [53]. *3DHOG* was implemented following [55], [11], i.e. computing 3D HOG track features for each tracked upper body in each video clip and then feed them into a SVM classifier with RBF kernel. These two baselines are two representative methods for action recognition in video and can produce state-of-the-art results. We also implement the *Explicit Model* [11] that describes the interactions between human and manipulated object by combining 3D HOG track features and

[1]Since the confusion matrix is a 24×24 table, it is omitted due to space constraint.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 6. Some examples from SYSU action dataset with our annotations.

DataSet	clutter	pose	noise	video num	in/outdoor
Gupta	small	standing	small	60	indoor
SYSU	large	standing & sitting	large	119	indoor & outdoor

TABLE III
COMPARING THE GUPTA AND SYSU ACTION DATASETS

Acc(mAP%)	IP [53]	3DHOG [55]	ExplicitModel[11]	Our model
Gupta	57.7(76.3)	61.5(74.9)	84.6(88.8)	**92.3**(**91.1**)
SYSU action	35.2(40.5)	39.4(50.0)	45.0(**53.0**)	**52.1**(51.7)

TABLE IV
RECOGNITION RESULTS ON THE GUPTA AND SYSU ACTION DATASETS IN RECOGNITION ACCURACY AND MEAN AVERAGE PRECISION (THE NUMBERS IN BRACKETS).

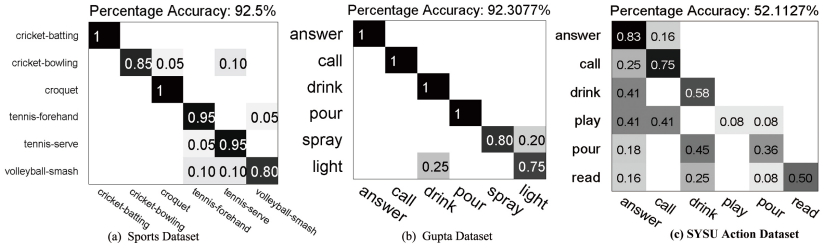
an interaction descriptor. This interaction descriptor consists of relative location, relative area and relative motion between human and the manipulated object. All the three interaction cues were computed by following [11] strictly. It differs from our approach in that it heavily relies on the detection of the manipulated objects and person which makes it sensitive to the detection errors caused by object or person tracking, while our approach can use the learned pose-object interaction exemplar to alleviate the influence of these errors. For the compared methods, we used the same classifier used in authors' papers. Note that the original model developed in [11] needs the users to manually annotate the initialised location of the manipulated object during both training and testing, which is less applicable in practice. Importantly, it needs more information, particularly during test, which put it an unfair advantage. For a fair comparison, in our experiment, the initial localisation of objects is done automatically based on the same detection results used in our model.

2) Results on the Gupta Dataset: The complete *Gupta* dataset consists of 60 video clips of 6 different HOI actions: drinking from a cup, spraying from a spray bottle, answering a phone call, making a phone call, pouring from a cup and lighting the flashlight. These actions were acted by 10 subjects. However, the set we can only manage to download from the authors' website consists of 54 video clips. During our experiment on this subset, we selected 28 videos for training and the rest for testing. Hence, we shall note that the results of our implementations are different from those reported in [11] as they were not obtained from the same set of data.

We report our comparison results on this dataset in Table IV. Our model achieves the best performance and outperforms other baselines by more than 7.7% in terms of accuracy. This demonstrates that our exemplar modelling is equally competitive for HOI action recognition in video. The confusion matrix of our model are presented in Figure 7(b). From the matrix, we can see that our model can distinguish actions of answering, calling, drinking and pouring correctly, which demonstrates that our model can effectively capture the interactions between human and the manipulated object. However, spraying and lighting are often confused with each other in our experiment. This is because that spraying and lighting often share similar object appearance and interaction cues.

3) Results on the SYSU action Dataset: We also test our proposal on a new dataset *(SYSU action)* consisting of 10 subjects who perform six interactions: answering a phone, calling a phone, playing with a phone, reading a book, drinking using a cup and pouring using a cup, with three different objects. This set consists of 119 video clips taken by an amateur camcorder. As shown in Figure 6, compared to the Gupta dataset, this dataset is more challenging because of the following reasons. 1) Different from the Gupta dataset, our SYSU dataset was collected in two different scenarios including indoor and outdoor, so that the lighting condition is more diverse. 2) There are more similar HOI action classes in the SYSU dataset. For example, three classes, answering a phone, making a phone call, and playing with a phone all involve the same type of object and similar human-object interactions. 3) In most classes the action was performed in two distinct poses: sitting and standing. This introduces much greater intra-class variations than those in the Gupta dataset. 4) In SYSU, the same set of objects appear in different action classes even though only one of them is manipulated by the actor. For example, when someone is answering phone, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology



(a) Sports Dataset (b) Gupta Dataset (c) SYSU Action Dataset

Fig. 7. Confusion matrices of the proposed method on the Sports image, Gupta and SYSU action video datasets.

cup (object involved in the pouring a drink action) and book (reading action) can also be founded in the vicinity (see row 1 of Figure 6). This reflects the real-world scenario (one cannot assumes that only one object can appear in a given scene), as well as poses great challenges to HOI action recognition as multiple objects will be detected and which one of them is interacting with the human in the scene needs to inferred. 5) The SYSU dataset contains about twice as many video clips as the Gupta dataset. We summarise the differences between the two datasets in Table V-B1. The SYSU dataset as well as our annotations will be released on our website soon.

We compare our method with the state-of-the-art alternatives on the cross-subject setting, where the examples of 4 subjects were selected for training and the rest of the examples were used as testing data. We tabulate the comparison results in Table IV. As can be seen, our model outperforms the baselines by 16.9%, 12.7% and 7.1% respectively in terms of accuracy, and most of the baselines in mAP values. From the table, we also observe that both our method and *ExplicitModel* outperform IP [53] and 3DHOG [55] with a big margin. This suggests that a human-object interactions modelling approach is more powerful than the conventional space-time interest points based approach for video based HOI action recognition. Note that the performance achieved by our model on this dataset $(52.1(51.7))$ is much lower than the results on the Gupta dataset $(92.3(91.1))$, because this new dataset is much more challenging.

The confusion matrices are shown in Figure 7(c). We observe that our model often confuses the interactions of pouring with drinking. This is because the two interactions are with the same manipulated object. It is observed that the main difference between them is that videos of drinking often contain a small number of key frames where cup appeared near the actor's mouse. We thus believe that the performance can be improved by integrating more elaborated key frame selection algorithm such as the work in [27].

C. Further Evaluations

The number of candidate exemplars and the detected locations are the two major parameters that affect our exemplar's performance. Here, we evaluate their influence. We also validate the effectiveness of our weighted pooling and matching model.

1) Effects of the Number of Candidate Exemplars: We study the effect of different numbers of exemplars S used when nominating atomic poses (Sec. III-C1). Figs. 8(a)-(d) show the performances of the proposed model on Sports,

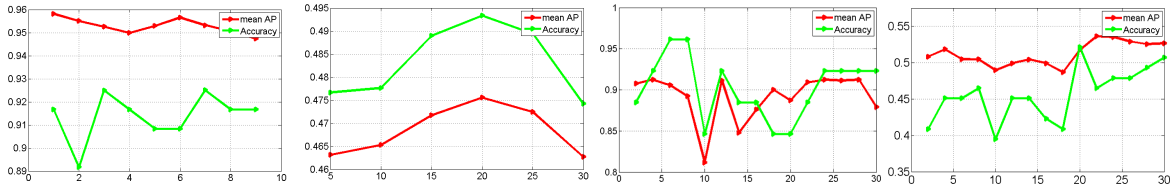
Accuracy(mAP)	Sports	PPMI	Gupta	SYSU action
Without Per.	92.5(95.25)	49.34(47.56)	92.31(91.14)	52.11(51.73)
With Per. (± 3)	91.67(95.74)	48.73(47.49)	92.31(90.00)	52.11(51.96)
With Per. (± 5)	89.17(95.42)	48.83 (47.56)	92.31(87.64)	50.70(50.89)
With Per.(± 10)	88.30(94.93)	48.67(47.29)	88.46(89.06)	50.70(51.34)

TABLE V
EFFECT OF PERTURBATIONS (3, 5 AND 10 PIXELS IN MAXIMUM) TO OUR MODEL IN CLASSIFICATION ACCURACY (%).

PPMI, Gupta and SYSU action datasets, respectively. The performance peaks when $S = 3$ on the Sports dataset, $S = 20$ on the PPMI dataset, $S = 6$ on the Gupta video dataset and $S = 20$ on the SYSU activion dataset, which is almost one fourth of the number of atomic poses learned for static image datasets and one twentieth of the atomic pose number for the video datasets. Overall, the accuracy first increases and then decreases as S increases. In comparison, the performance of the model is more sensitive to S on the PPMI and SYSU activion datasets. This is because PPMI and SYSU action datasets have more variations of human pose. Hence better performances on PPMI and SYSU can be obtained for a larger S, as more candidate exemplars are needed to describe the spatial pose-object interaction in an image. In the future work, an optimisation technique may be necessary to estimate it automatically.

2) Influence of Perturbation in Detections: We evaluate the robustness of our model given errors in person/object detection(tracking). 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 V, tabulating both accuracy and mAP results where mAP results are in the brackets. The results show that the performance drops only slightly by 1%$\sim$5% in accuracy and less than 2% 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$8% more than the compared methods in accuracy), and outperforms SPM and Grouplet on PPMI by 5%$\sim$7% in mAP, and still performs comparably with Yao's method. A similar conclusion can be drawn for the SYSU action dataset. Note that no performance drop can be observed on the Gupta set in terms of accuracy when $p = 3$ and $p = 5$. It might be because the exploitation of multiple frame information alleviates the effect of perturbation in each frame. These results suggest that our model is robust against errors

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2397200, IEEE Transactions on Circuits and Systems for Video Technology



(a) Sports (b) PPMI (c) Gupta (d) SYSU Action

Fig. 8. Performance of our models given different numbers of candidate exemplar (i.e. S) on Sports, PPMI, Gupta and SYSU action dataset. The X-axis and Y-axis indicate selected exemplar number and the performance (accuracy) respectively.

Accuracy(mAP)	Gupta	SYSU action
Proposed Weighted Average (Eq.(9))	92.3(91.1)	52.1(51.7)
Traditional Average	92.3(90.1)	47.9(50.0)
Max	88.5(90.5)	50.0(43.7)

TABLE VI
COMPARISON RESULTS OF DIFFERENT POOLING METHODS(%).

Accuracy	Sports	PPMI	Gupta	SYSU action
Matching Model	**92.5**	**49.3**	**92.3**	**52.1**
LSVM	91.7	48.4	80.8	45.2
Naive Weights	87.5	43.1	76.9	40.9
MKL [47]	89.1	49.1	88.5	47.9
RankSVM	89.1	43.3	73.1	33.8

TABLE VII
COMPARING OUR MATCHING MODEL AGAINST THE ALTERNATIVES. THE RESULTS ARE IN CLASSIFICATION ACCURACY (%).

in person/object detection.

3) Effect of the weighted pooling in video frames: As discussed in Sec. IV, we employ a weighted pooling method (Eq. (9)) to integrate all the responses calculated from video frames. As can be observed from Table VI, the performance would drop when we replace our weighted pooling with traditional average pooling and max pooling methods. The results indicate that the implicit frame selection by our weighted pooling method is useful for alleviating the effect of irrelevant or noisy frames during video HOI modelling.

4) Effectiveness of the matching model: Our matching model attempts to learn a set of weights to measure the reliability of each component in the HOI descriptor. Here, we show the advantage of the proposed matching model over the traditional learning methods such as multiple kernel learning [47], linear SVM, Ranking SVM and a naive weighted method by simply averaging over all the components. We present the results in Table VII. It can be observed that our matching model can achieve the best classification accuracy on all four datasets. The margin is particularly big on the two video datasets. This result demonstrates that the proposed direct ranking method can learn more robust and discriminant information about the relative weighting of different components in the HOI descriptor. We also observe that RankSVM performs worse than the baseline Naive Weights on the two video datasets. The poor performance of RankSVM on the two video datasets can be largely explained by the fact that those action classes have a much greater degree of inter-class variations due to more temporal, lighting and pose changes than the ones on still image datasets, so a single set of weighting for all classes are no longer appropriate. Moreover, RankSVM employs relative comparisons of all the sample pairs to train the model, which may suffer from the over-fitting problem with the two relatively smaller size video datasets.

5) Effectiveness of individual HOI descriptor components: In Table VIII, we evaluated the contribution of each component in our full HOI descriptor by removing one component from the full descriptor. The results show that all the components have contributed positively towards the superior results of our method. Furthermore, we can make the following observations: (1) Among the four components of the HOI descriptor (scene context, object detection vector, pose appearance feature, and the novel spatial exemplar feature), overall on the four datasets, the proposed spatial exemplar-based HOI feature has the biggest contributions, as on average the performance drops the most when this feature is removed. (2) The scene context is important in the two still image datasets (Sports and PPMI). This is because the background of the HOI actions in each class in these two sets is often similar, and dissimilar to that of other classes. Therefore scene context becomes a very effective cue. For instance, in the Sports dataset, the action 'croquet' often occurs outdoor on green grass, whilst the action 'volleyball smash' often occurs in an indoor volleyball court. However, for video HOI actions, scene context is less important because most action classes feature sequences were captured in the same scene. For these dataset, our results show that the object detection and the spatial interaction exemplar cues have a stronger influence on the performance of our video-based HOI recognition method. (3) The recognition accuracy decreases on the SYSU dataset (52.1 to 38.0) to a greater extent than that on the Gupta dataset (92.3 to 84.6) when removing the spatial interaction. This is because in the SYSU dataset, the same set of objects (phone, cup and book) appears in most HOI action instances regardless of the action class. The spatial relations between human subject and the manipulated object thus become even more important for distinguishing different actions.

VI. CONCLUSION AND FUTURE WORK

We have proposed to represent human-object interactions using a set of spatial pose-object interaction exemplars and formed a new HOI descriptor consisting four parts, where the weights for each part are learned by a ranking model. A key characteristic of our exemplar-based approach is that it models the mutual structure between human and object in a probabilistic way, so as to avert explicit human pose

Method	Sports	PPMI	Gupta	SYSU action
Full	92.5(95.3)	49.3(47.6)	92.3(91.1)	52.1(51.7)
Without Context	88.3(92.5)	43.3(41.2)	84.6(83.1)	46.5(48.9)
Without Object	88.3(94.8)	45.6(44.0)	80.8(81.3)	43.7(42.2)
Without Pose	88.3(94.1)	48.0(46.8)	88.5(86.0)	49.3(51.1)
Without Spatial Interaction	86.7(92.1)	44.6(43.4)	84.6(83.7)	38.0(41.3)

TABLE VIII
ACTION RECOGNITION RATES (%) MEASURED BY BOTH ACCURACY AND (MAP) USING THE HOI DESCRIPTOR WITH VARIOUS FEATURES REMOVED IN THE PROPOSED MODEL

estimation and alleviate the effects of imperfect detection of object and human. The proposed exemplar modelling is also able to reduce irrelevant and noisy frames during HOI modelling on videos. Our experimental results suggest that our exemplar approach is able to outperform most existing related HOI techniques in both still images and video frames. On-going work focuses on 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 the future, we consider exploring the use of large scale data mined from the internet for learning exemplars. Note that our exemplars are obtained based on a large amount of manual annotations during training, which may be expensive for dealing much more larger scale data. So exploring a weakly supervised or even unsupervised method for learning a set of pose-object interaction exemplars is another direction of our future work.

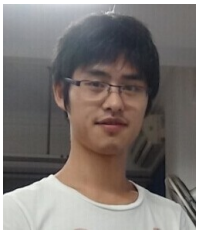
ACKNOWLEDGMENT

The SYSU datasets and features employed in the proposed method would be available at: http://sist.sysu.edu.cn/%7ezhwshi/HOI.html

REFERENCES

[1] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interactions activities," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[2] A. Prest, C. Schimid, and J. Malik, "Weakly supervised learning of interactions between humans and objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 601–614, 2012.
[3] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," *In Workshop on Structured Models in Computer Vision*, 2010.
[4] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[5] ——, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, 2012.
[6] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[7] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[8] V. Delaitre, J. Sivic, I. Laptev *et al.*, "Learning person-object interactions for action recognition in still images," in *Annual Conference on Neural Information Processing Systems*, 2011.
[9] A. Gupta, A. Kembhavi, and L. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.
[10] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
[11] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 835–848, 2013.
[12] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.
[13] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.
[14] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *European Conference on Computer Vision*. Springer, 2012, pp. 158–172.
[15] B. Yao and L. Fei-Fei, "Action recognition with exemplar based 2.5 d graph matching," in *European Conference on Computer Vision*, 2012.
[16] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[17] B. Yao, Z. Liu, X. Nie, and S. Zhu, "Animated pose templates for modelling and detecting human actions," 2013.
[18] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.
[19] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," *International Conference on Computer Vision*, 2009.
[20] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations." in *British Machine Vision Conference*, vol. 2, no. 5, 2010, p. 7.
[21] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," *International Conference on Computer Vision*, 2007.
[22] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
[23] G. Sharma, F. Jurie, C. Schmid *et al.*, "Expanded parts model for human attribute and action recognition in still images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[24] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg, "Coloring action recognition in still images," *International Journal of Computer Vision*, pp. 1–17, 2013.
[25] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff, "Learning actions from the web," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 995–1002.
[26] C. Thurau and V. Hlavác, "Pose primitive based human action recognition in videos or still images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
[27] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.
[28] R. Filipovych and E. Ribeiro, "Recognizing primitive interactions by exploring actor-object states," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–7.
[29] H. Kjellström, J. Romero, and D. Kragić, "Visual object-action recognition: Inferring object affordances from human demonstration," *Computer Vision and Image Understanding*, vol. 115, no. 1, pp. 81–90, 2011.
[30] J. Arpit, G. Abhinav, R. Mikel, and D. L. S, "Representing videos using mid-level discriminative patches," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[31] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1194–1201.
[32] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *IEEE International Conference on Computer Vision*, 2009, pp. 104–111.
[33] H. S. Koppula and A. Saxena, "Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation."
[34] P. Wei, Y. Zhao, N. Zheng, and Z. Song-Chun, "Modeling 4d human-object interactions for event and object recognition," in *IEEE International Conference on Computer Vision*, 2013.
[35] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3d human action recognition," in *Human Action Recognition with Depth Cameras*. Springer, 2014, pp. 11–40.

[36] J. H. J. Xiao, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[37] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," *International Conference on Computer Vision*, 2011.
[38] G. Mori and J. Malik, "Recovering 3d human body configurations using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1052–1062, 2006.
[39] W. Brendel and S. Todorovic, "Activities as time series of human postures," in *European Conference on Computer Vision*. Springer, 2010, pp. 721–734.
[40] B. Yao, J. Ma, and L. Fei-Fei, "Discovering object functionality," in *IEEE International Conference on Computer Vision*, 2013.
[41] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang, "Recognising human-object interaction via exemplar based modelling," in *IEEE International Conference on Computer Vision*, 2013.
[42] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.
[43] J. Wu, W.-S. Zheng, and J.-H. Lai, "Approximate kernel competitive learning," *Neural Networks*, vol. 63, no. 1, pp. 117–132, 2015.
[44] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," *International Conference on Computer Vision (ICCV)*, 2007.
[45] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, 2012.
[46] W.-S. Zheng, S. Gong, and T. Xiang, "Quantifying and transferring contextual information in object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 762–777, 2012.
[47] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1065–1072.
[48] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.
[49] J. J. Wang, J. C. Yang, K. Yu, F. J. Lv, T. Huang, and Y. H. Gong, "Locality-constrained linear coding for image classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[50] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2013.
[51] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
[52] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
[53] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1948–1955.
[54] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549–1562, 2012.
[55] A. Kläser, M. Marszałek, C. Schmid, and A. Zisserman, "Human focused action localization in video," in *Trends and Topics in Computer Vision*. Springer, 2012, pp. 219–233.

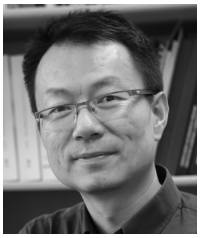


Jian-Fang Hu received the B.S. and M.S. degrees from the School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, China, in 2010 and 2012, respectively, and he is currently pursuing the Ph.D. degree with the Department of Applied Mathematics. His current research interests include computer vision and applied machine learning, including human-object interaction modeling, 3D face modeling, and RGB-D activity recognition.

Wei-Shi Zheng is currently an Associate Professor of Sun Yat-sen University. He joined the University in 2011. His research direction is machine vision and intelligence learning. His current interests are in object association and activity analysis in computer vision, particularly focusing on person re-identification and activity recognition for visual surveillance. He has now published more than 60 papers, including more than 40 publications in main journals (TPAMI,TNN,TIP,TSMC-B,PR) and top conferences (ICCV, CVPR,IJCAI,AAAI). He has joined the organisation of four tutorial presentations in ACCV 2012, ICPR 2012, ICCV 2013 and CVPR 2015 along with other colleagues. He has been awarded the new star of science and technology of Guangzhou in 2012 and Guangdong natural science funds for distinguished young scholars in 2013.

Jianhuang Lai received the M.Sc. degree in applied mathematics in 1989 and the Ph.D. degree in mathematics in 1999 from Yat-sen University, China. He joined Sun Yat-sen University in 1989 as an Assistant Professor, where he is currently a Professor with the Department of Automation of School of Information Science and Technology, and Dean of School of Information Science and Technology. His current research interests include the areas of digital image processing, pattern recognition, multimedia communication, wavelet and its applications. He has published more than 100 scientific papers in the international journals and conferences on image processing and pattern recognition, such as IEEE TPAMI, IEEE TKDE, IEEE TNN, IEEE TIP, IEEE TSMC (Part B), PR, ICCV, CVPR and ICDM. Prof. Lai serves as a Standing Member of the Image and Graphics Association of China, and also serves as a Standing Director of the Image and Graphics Association of Guangdong. He is a senior member of the IEEE.

Shaogang Gong is Professor of Visual Computation at Queen Mary University of London, a Fellow of the Institution of Electrical Engineers and a Fellow of the British Computer Society. He received his D.Phil in computer vision from Keble College, Oxford University in 1989. His research interests include computer vision, machine learning and video analysis.



Tao Xiang received the PhD degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published over 100 papers in international journals and conferences and co-authored a book, Visual Analysis of Behaviour: From Pixels to Semantics.