Neurocomputing 321 (2018) 216–226

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Understanding human activities in videos: A joint action and interaction learning approach

Zhenhua Wang[a,*], Jiali Jin[a], Tong Liu[a], Sheng Liu[a], Jianhua Zhang[a], Shengyong Chen[a], Zhen Zhang[b], Dongyan Guo[a], Zhanpeng Shao[a]

[a] College of Computer Science and Technology, Zhejiang University of Technology, 288 Liuhe Road, Xihu District, Hangzhou 310023, PR China
[b] National University of Singapore, 21 LowerKent Ridge Road, Singapore

ARTICLE INFO

Article history:
Received 18 May 2018
Revised 1 September 2018
Accepted 12 September 2018
Available online 21 September 2018

Communicated by Yongdong Zhang

Keywords:
Structured prediction
Action recognition
Activity understanding

ABSTRACT

In video surveillance with multiple people, human interactions and their action categories preserve strong correlations, and the identification of interaction configuration is of significant importance to the success of action recognition task. Interactions are typically estimated using heuristics or treated as latent variables. However, the former usually introduces incorrect interaction configuration while the latter amounts to solve challenging optimization problems. Here we address these problems systematically by proposing a novel structured learning framework which enables the joint prediction of actions and interactions. To this end, both the features learned via deep nets and human interaction context are leveraged to encode the correlations among actions and pairwise interactions in a structured model, and all model parameters are trained via a large-margin framework. To solve the associated inference problem, we present two optimization algorithms, one is alternating search and the other is belief propagation. Experiments on both synthetic and real dataset demonstrate the strength of the proposed approach.

1. Introduction

Understanding human activities in videos is a fundamental problem in computer vision, and is essential to a number of interesting applications such as sports analysis, video retrieval and video surveillance. Recent progress on deep neural networks boosts the recognition performance on many tasks significantly [1–4], including action recognition [5–10]. However, these approaches predict each video a single action category, hence are inherently ineffective to obtain a fine-grained understanding of activities that contain multiple people with interactions, *e.g.*handshake, push, kick, meeting, football play *etc*. In order to support a deeper understanding of human activities in such videos, it is essential to recognize both the action of each individual and their interactions, which are beyond the scope of most existing work.

Probabilistic graphical representations have widely been utilized in many vision tasks [11–15]. Undirected graphical models, typically conditional random fields (CRFs) combine both local feature representations [13,16–18] and global interaction contexts [17,17,19–22] to represent human interactions. Such representation can be viewed as an effective fusion of rich features, which helps to boost the recognition performance evidently [13].

Assuming the image contains a number of n people. The term *human interaction configuration* (HIC) is a symmetric matrix $\mathbf{E} \in \{0, 1\}^{n \times n}$ with $\mathbf{E}_{i,j} = 1 \ \ \forall i, j \in \{1, 2, \ldots, n\}, i \neq j$ representing person i and j are interacting with each other (like handshaking, hugging, chasing and fighting), and $\mathbf{E}_{i,j} = 0$ indicates the associated people are not interacting. In addition, $\mathbf{E}_{i,i} = 0 \ \ \forall i \in \{1, 2, \ldots, n\}$. Within a CRF model, the associated graph encodes HIC via edges, where each edge represents that the associated people are interacting with each other, while the vacancy of an edge indicates that the associated individuals have no interaction. An important problem, though has long been neglected within the community of action recognition, is that the topology (or structure) of CRF graphs is typically predefined using domain knowledge or heuristics, *e.g.*distances between persons [17], complete graph [12] or hierarchical structures [11], which can introduce undesirable or irrelevant HIC, consequently misguiding the training of CRF parameters and deteriorating the recognition performance.

To obtain proper CRF graphs for action recognition, another branch is to infer CRF graphs (*i.e.* HIC) directly from data. Since HIC are typically heterogeneous across different activities and instances, traditional graph learning approach (see Chapter 18 in [23]) is not applicable. To tackle this, Lan et al. [19] and our past work [20] use the latent structured SVM framework [24] to jointly

* Corresponding author.
E-mail address: zhhwang@zjut.edu.cn (Z. Wang).

https://doi.org/10.1016/j.neucom.2018.09.031
0925-2312/

model human interactions and actions, within which interactions are treated as latent variables. The problem with such approaches are 1) the learned HIC can be un-interpretable due to the lack of supervision on interaction during training; 2) the associated training problem is non-convex and it is challenging to train good models.

In this paper, our task is to obtain a frame-wise prediction of the action categories of all individuals and their pairwise HIC. To this end, we propose a novel CRF model which enables the estimation of human actions and interactions in videos jointly and simultaneously without introducing latent variables, *i.e.* we learn both actions and interactions in a supervised manner, which outperforms the baseline methods by a large margin, while is at least competitive in comparison with the state-of-the-arts. This work is an extension of our joint interaction and action learning approach [25]. The extensions include 1) deriving a new algorithm which solves the inference problem in the joint learning framework and comparing its effectiveness against the original inference in [25]; 2) evaluating the proposed approach on one additional dataset; 3) studying the effect of human body detection to performance of action recognition. The closest approaches to ours are the structured learning approaches [13,26]. Work [26] learns interactions and actions jointly in a supervised manner without using latent variables as well. The main difference between this approach and ours is that we propose a new potential function which combines handcrafted descriptors, learned features and contextual information to depict human interactions, which in turn delivers superior recognition results. Moreover, to solve the relevant inference problem, they resort to exhaustive search, while we provide two efficient optimization algorithms which scale to large problems with many variables. Our recent work [13] casts the CRF graph learning into a problem classifying *interactions* and *non-interactions* with linear support vector machine (SVM) classifiers [13]. Then the classifications are taken to determine the graphs within the proposed spatial-temporal CRFs. Hence in [13] the interactions and actions are learned separately rather than jointly.

2. Existing approach

We now review two important approaches for human action recognition, where interactions are either constructed based on heuristic rules or inferred from data. These approaches are taken to compare against our method in Section 3.

2.1. CRF with heuristic HIC

First we introduce some notations which will be used throughout this paper. Let $G=(V,E)$ be a graph, where the node set V representing all individuals within a frame, and the edge set E representing their HIC. For instance, the connection $e_{ij}\in E$ indicates that the associated person i and j have interaction with each other, while the absent of an edge between node s and t means that person s and t are not interacting.

CRF with heuristic HIC is usually taken as the baseline for human interaction modeling [19]. This approach determines E according to some heuristics (*e.g.*interacting people are close in distance), based on which distant people are ruled out from E using a predesignated threshold. Let $\mathbf{I}$ denote an arbitrary frame in a video. Let $a_i\in\mathbb{A}$ be the action label of person i, and $\mathbf{a}=[a_i]_{i=1,\dots,n}$ be a vector including the action labels of n persons. The energy function of this CRF representation is

$$g(\mathbf{a};\theta)=\sum_{i\in V}\sum_{s\in\mathbb{A}}\theta_{i;s}\mathbb{1}_s(a_i)+\sum_{(u,v)\in E}\sum_{(s,t)\in\mathbb{A}^2}\theta_{u,v;s,t}\mathbb{1}_{s,t}(a_u,a_v), \quad (1)$$

where $\theta_{i;s}$ is the unary energy when node i takes a label $s\in\mathbb{A}$, $\theta_{u,v;\,s,t}$ is the pairwise energy when a pair of interacting people (u, v) takes $(s,t)\in\mathbb{A}^2$. These energies are typically computed with the formulation below:

$$\theta_{i;s}=\mathbf{w}_u^\top\phi_i(\mathbf{I},s), \quad (2)$$

$$\theta_{u,v;s,t}=\mathbf{w}_b^\top\phi_{uv}(\mathbf{I},s,t), \quad (3)$$

where $\mathbf{w}=[\mathbf{w}_u;\mathbf{w}_b]$ are the model parameters to be learned from training samples, and ϕ_i, ϕ_{uv} are the so-called joint feature vectors (see Section 3 for details). A joint feature vector is a combination of observed information and labels, and a dot-product between the join feature and the model parameters gives an evaluation of the compatibility between labels and observation.

Within (1), $\mathbb{1}_s$, $\mathbb{1}_{s,t}$ are indicator functions defined as

$$\mathbb{1}_s(a_i)=\begin{cases}1 & a_i=s,\\ 0 & \text{otherwise.}\end{cases} \quad (4)$$

$$\mathbb{1}_{s,t}(a_u,a_v)=\begin{cases}1 & a_u=s \ \text{ and } \ a_v=t,\\ 0 & \text{otherwise.}\end{cases} \quad (5)$$

A good point of CRF with heuristic interactions is that the associated training problem of (1) is convex, hence one can learn $\mathbf{w}$ with maximum a likelihood estimation (MLE) or structured support vector machine (SSVM), during which the related inference problems (*i.e.* find the most probable $\mathbf{a}$) are approximately solved via a belief propagation-style algorithm.

2.2. CRF with latent HIC

Lan et al. [19] proposed a discriminant model to estimate actions and interactions jointly from data. Within this model, HICs are treated as latent variables, *i.e.* the related training process only penalizes incorrect action predictions. The associated energy function is

$$h(\mathbf{a},\mathbf{e};\theta)=\sum_{i\in V}\sum_{s\in\mathbb{A}}\theta_{i;s}\mathbb{1}_s(a_i)+\sum_{u,v\in V,u<v}\ \sum_{(s,t)\in\mathbb{A}^2}e_{uv}\theta_{u,v;s,t}\mathbb{1}_{s,t}(a_u,a_v), \quad (6)$$

which is identical to (1) except for multiplying each pairwise energy $\theta_{u,\,v}$ by an introduced binary latent variables $e_{uv}\in\{0,1\}$. Here $e_{uv}=1$ indicates that the instantaneous interaction between $(u,\,v)$ takes place, and vice versa.

Let $\{(\mathbf{I}^k,\mathbf{a}^k)\}_{k=1}^N$ be a set containing N training examples. The related training problem (using a max-margin formulation) is

$$\begin{aligned}&\min\ \frac{1}{2}\|\mathbf{w}\|^2+\frac{C}{N}\sum_{l=1}^{N}\xi_l^2\\ &\text{s.t.}\ \ \max_{\hat{\mathbf{e}}^l}h(\hat{\mathbf{a}}^l,\hat{\mathbf{e}}^l;\theta)-\max_{\mathbf{e}^l}h(\mathbf{a}^l,\mathbf{e}^l;\theta)\geq\Delta(\mathbf{a}^l,\hat{\mathbf{a}}^l)-\xi_l,\\ &\qquad\hat{\mathbf{a}}^l\neq\mathbf{a}^l,\xi_l\geq 0\ \ \forall l,\end{aligned} \quad (7)$$

where C is the tradeoff between model complexity and the empirical risk, ξ is the slack variable and Δ is the label cost defined by

$$\Delta(\mathbf{a},\hat{\mathbf{a}})=\frac{1}{n}\sum_{i=1}^{n}\delta(a_i\neq\hat{a}_i). \quad (8)$$

Above n is the number of people in the image, δ is an indicator function that gives 1 if the testing condition is true, and outputs 0 otherwise. Since problem (7) is non-convex, one can only get suboptimal solutions of $\mathbf{w}$ with approximations like that proposed in [27].

During training and prediction, one need to repeatedly solve the following inference problem:

$$\min_{\mathbf{a},\mathbf{e}}\ \ h(\mathbf{a},\mathbf{e};\theta)\ \ \text{s.t. degree}(i)\leq d,\forall i\in V. \quad (9)$$

Here the constraints ensure the estimation of sparse CRF graphs with node degree less than a constant d. The problem is NP hard and Lan et al. [19] solve it via an iterative procedure. In each iteration, the algorithm alternatingly performs two steps: 1) fixing action variables and search in the interaction space (solving an integer programming problem); 2) fixing interactions and search in the action space (solving a CRF inference problem with known graphs).

Rather than using alternating search, Wang et al. [20] relaxed the original optimization (9) into a bilinear programming problem (BLP) by replacing the indicator functions $\mathbb{1}(\cdot) \in \{0, 1\}$ with continuous variables $\mu \in [0, 1]$, thus (9) becomes

$$\min_{\mu,\mathbf{e}} \sum_{i\in V}\sum_{s\in\mathbb{A}} \theta_{i;s}\mu_{i;s} + \sum_{i,j\in V, i<j}\sum_{(s,t)\in\mathbb{A}^2} e_{ij}\theta_{i,j;s,t}\mu_{i,j;s,t}$$
$$\text{s.t. } \text{degree}(i) \le d,\ (\mu,\mathbf{e}) \in \mathcal{O}, \forall i \in V. \tag{10}$$

Above $\mathcal{O}$ denotes a space given by

$$\mathcal{O} = \left\{ \mu, \mathbf{e} \;\middle|\; \begin{array}{l} \mu_{i,j;s,t}, e_{ij} \in [0,1], \forall i < j, s, t, \\ \sum_s \mu_{i;s} = 1, \forall i \in V, s, \\ \sum_s \mu_{i,j;s,t} = \mu_{j;t}, \forall i < j, t, \\ \sum_t \mu_{i,j;s,t} = \mu_{i;s}, \forall i < j, s. \end{array} \right\} \tag{11}$$

In [20], the BLP problem is solved via branch and bound strategy, with the bounds computed by solving a linear programming (LP) relaxation of BLP:

$$\begin{aligned} \min_{\mu,\lambda,\mathbf{e}} & \sum_{i\in V}\sum_{s\in\mathbb{A}} \mu_{i;s}\theta_{i;s} + \sum_{i,j\in V, i<j}\sum_{(s,t)\in\mathbb{A}^2} \theta_{i,j;s,t}\lambda_{i,j;s,t}, \\ \text{s.t. } & \lambda^l = \max\{\mu^l_{i,j;s,t}e_{ij} + e^l_{ij}\mu_{i,j;s,t} - \mu^l_{i,j;s,t}e^l_{ij}, \mu^u_{i,j;s,t}e_{ij} \\ & \quad + e^u_{ij}\mu_{i,j;s,t} - \mu^u_{i,j;s,t}e^u_{ij}\}, \\ & \lambda^u = \min\{\mu^u_{i,j;s,t}e_{ij} + e^l_{ij}\mu_{i,j;s,t} - \mu^u_{i,j;s,t}e^l_{ij}, \mu^l_{i,j;s,t}e_{ij} \\ & \quad + e^u_{ij}\mu_{i,j;s,t} - \mu^l_{i,j;s,t}e^u_{ij}\}, \\ & \lambda^l \le \lambda_{i,j;s,t} \le \lambda^u,\ (\mu,\mathbf{e}) \in \mathcal{O}, \text{degree}(i) \le d, \forall i < j, s, t. \end{aligned} \tag{12}$$

Here $[e^l_{ij}, e^u_{ij}]$ and $[\mu^l_{i,j;s,t}, \mu^u_{i,j;s,t}]$ denote any sub-regions of $e_{ij} \in [0, 1]$ and $\mu_{i,j;s,t} \in [0, 1]$ obtained by the branching operation.

3. Joint action and interaction learning

Given a frame, our joint-learning formulation predicts both the action labelling $\mathbf{a}$ and the HIC via solving

$$\min_{\mathbf{a},\mathbf{e}} \quad f_0(\mathbf{a}, \mathbf{e}; \theta) \quad \text{s.t.} \quad a_i \in \mathbb{A}, e_{ij} \in \{0, 1\}\ \forall i < j, \tag{13}$$

where

$$\begin{aligned} f_0(\mathbf{a},\mathbf{e};\theta) = & \sum_{i<j}\sum_{z\in\{0,1\}} \theta_{i,j;z}\mathbb{1}_z(e_{ij}) + \theta_0\|\mathbf{e}\|_1 + \sum_{i\in V}\sum_{s\in\mathbb{A}} \theta_{i;s}\mathbb{1}_s(a_i) \\ & + \sum_{i<j}\sum_{(s,t)\in\mathbb{A}^2} e_{ij}\theta_{i,j;s,t}\mathbb{1}_{s,t}(a_i, a_j), \end{aligned} \tag{14}$$

where $\mathbb{1}_z(e_{ij})$ is another indicator which gives 1 if $e_{ij} = z$ (and 0 otherwise). In comparison with the CRF formulation (1) and (6), the main differences include: 1) The formulation (6) only learns action labels from data while treating the HIC as latent variables (our method learns both labels and interactions jointly and explicitly), which results in a challenging non-convex optimization problem (note our training problem is convex). To solve the associated non-convex optimization, it requires a careful initialization of the model parameters to get a good performance. 2) Our energy function (14) contains two novel terms, one is used to represent whether an interaction exists based on image cues, and the other is taken to ensure the sparse estimation of interactions. We show in Section 4 that our joint learning formulation helps to improve recognition performance on both actions and interactions. Terms of our formulation (14) are detailed below.

3.1. Unary energy

The unary energy $\theta_{i;s}$ is defined by

$$\theta_{i;s} = -p_{i;s}\mathbf{w}_u^\top \mathbf{1}_u(s), \tag{15}$$

where $\mathbf{w}_u$ is the weight of the unary energy, $\mathbf{1}_u(s) \in \{0, 1\}^{|\mathbb{A}|}$ is an indicator vector that takes 1 at the sth position and 0 elsewhere, $p_{i;s}$ is the discriminant score when assigning label s to person i. To compute the discriminant scores, we train support vector machine (SVM) classifiers using a concatenation of the deep representations learned with two-stream neural networks [5], the histogram of gradients (HOG) features and the histogram of optical flow (HOF) features extracted from human body areas. To capture motion more accurately, we extract warped optical flow using the method described in [28]. Note that combining deep features and hand-engineered descriptors is a widely applied technique in action recognition [5,10], as deep feature alone typically fails to evidently outperform fine-grained hand-designed features for this task. Indeed, the concatenated feature outperforms its each standalone component significantly, see Section 4 for details.

3.2. Pairwise energy

The pairwise energy is given by

$$\theta_{i,j;s,t} = -\mathbf{w}_c^\top \mathbf{1}_c(s, t; l_{i,j}; p_i, p_j). \tag{16}$$

Here $\mathbf{w}_c$ measures the compatibility between the action labels s and t of two interacting persons, $\mathbf{1}_c(s, t; l_{i,j}; p_i, p_j) \in \{0, 1\}^{|\mathbb{L}\times\mathbb{P}\times\mathbb{A}^2|}$ is an indicator vector that takes 1 at the position indexed by $(l_{i,j}, p_i, p_j, s, t)$, and it takes 0 elsewhere. Here $l_{i,j} \in \mathbb{L}$ denotes the relative distance of person i to person j, which is computed by

$$l_{i,j} = \left\lceil h_{i,j}/\bar{h} \right\rceil. \tag{17}$$

Let r_i, r_j be the widths of the bounding boxes of person i, j. Let $d_{i,j}$ be the Euclidean distance between the centers of these two bounding boxes. We compute $h_{i,j}$ by

$$h_{i,j} = \frac{2d_{i,j}}{r_i + r_j}. \tag{18}$$

In Eq. (17), $\bar{h}$ denotes the mean h value of all interacting pairs (i, j) selected from training samples.

As typically done in action recognition [20,26], we categorize human head orientations into five classes, which are $\{profile-left, profile-right, frontal-left, frontal-right, backwards\}$. To determine the head orientation of each person, we train a linear large-margin classifier using both hand-crafted feature and the feature learned with a convolutional neural network [29]. We provide head-orientation-classification results in Fig. 5. The reason to use head orientations, instead of human-body orientations, is because estimating head orientations is easier than recognizing body orientations as human bodies are more likely to occlude with each other than heads.

3.3. Interaction energy

This energy represents the negative confidence with respect to the existence of interaction among people, which is defined by

$$\theta_{i,j;z} = -\{p_{i,j;z}\mathbf{w}_\tau^\top \mathbf{1}_\tau(z) + \mathbf{w}_r^\top \mathbf{1}_r(l_{i,j}; p_i, p_j; z)\}. \tag{19}$$

The parameter $\mathbf{w}_\tau$ weights the score $p_{i,j;z}$, representing the possibility of two persons (i, j) having interaction (when $z = 1$) or not (when $z = 0$). To calculate the score $p_{i,j;z}$, again we train SVM classifier using features extracted from areas that enclose bounding boxes of i and j. , and the feature representation here is similar to the features used to compute $p_{i;s}$ within the unary energy.

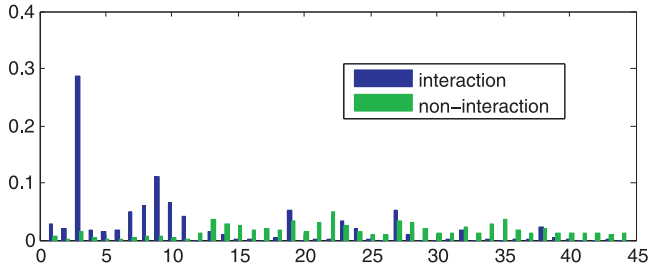


Fig. 1. Histograms of *relative position-head orientation* for interacting and non-interacting examples. Here each histogram is normalized and the bins with values smaller than 0.01 are discarded. Clearly the distributions of interacting and non-interacting examples are distinct. Hence relative human-body positions and head orientations are important cues to discriminate interacting and non-interacting people.

$\mathbf{1}_\tau(z) \in \{0, 1\}^2$ is an indicator vector which takes 1 at its zth position, and takes 0 elsewhere.

The parameter $\mathbf{w}_r$ evaluates the compatibility among the HIC (encoded by z), the relative distance l_{ij} and the head orientations p_i, p_j. As that shown in [20,26], head orientations are helpful to identify interacting people. This is because that interacting persons are likely to face each other (in activities like *e.g.*handshaking, hugging, kissing, punching). Hence we can leverage such contextual cues to promote the identification of HIC.

To illustrate the function of contextual information for interaction recognition, Fig. 1 visualizes the distributions with respect to these contextual cues for both interacting and non-interacting classes. One can find that these two classes can be well-separated with such knowledge.

3.4. Regularization term

The regularization term we introduced in (13) is

$$\theta_0 \|\mathbf{e}\|_1 = \sum_{i<j} e_{ij} w_s, \tag{20}$$

where $\theta_0 = w_s \in R$. When relaxing $e \in \{0, 1\}$ to $e \in [0, 1]$, this regularization term follows the definition of L_1 norm, which ensures that the learned HIC is sparse as one person is more likely to interact with one another rather than many.

3.5. Training

Assuming that we have a set of training instances $\{(\mathbf{I}^k, \mathbf{a}^k, G^k)\}_{k=1}^N$ with $G^k = (V^k, E^k)$. Here the edge set E^k represents the real HIC (*i.e.* the groundtruth) among people within the kth training data, and $\mathbf{e}^k$ is the vector form of E^k. We train all model parameters $\mathbf{w} = [\mathbf{w}_u, \mathbf{w}_c, \mathbf{w}_\tau, \mathbf{w}_r, w_s]$ with the following max-margin-style formulation:

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{l=1}^{n} \xi_l^2 \\ &\text{s.t. } f_0(\hat{\mathbf{a}}^l, \hat{\mathbf{e}}^l; \theta) - f_0(\mathbf{a}^l, \mathbf{e}^l; \theta) \geq \Delta(\mathbf{a}^l, \mathbf{e}^l, \hat{\mathbf{a}}^l, \hat{\mathbf{e}}^l) \\ &\quad - \xi_l, \forall l\ \hat{\mathbf{a}}^l \neq \mathbf{a}^l, \hat{\mathbf{e}}^l \neq \mathbf{e}^l, \xi_l \geq 0. \end{aligned} \tag{21}$$

The label cost, *i.e.* the penalty for incorrect predictions is

$$\Delta(\mathbf{a}, \mathbf{e}, \hat{\mathbf{a}}, \hat{\mathbf{e}}) = \frac{1}{m}\sum_{k=1}^{m} \delta(a_k \neq \hat{a}_k) + \frac{2}{m(m-1)}\sum_{i<j} \delta(e_{ij} \neq \hat{e}_{ij}), \tag{22}$$

where $\delta(\cdot)$ is the indicator function which gives 1 if the testing condition is true, and outputs 0 otherwise. m denotes the number of people. Since problem (21) is a convex optimization, we solve it using the cutting plane method [30] for structured prediction.

3.6. The inference problem

Now we discuss how to solve the challenging inference problem (13) with fixed θ parameters. Since the problem is NP hard, it is impossible to obtain global solutions in general. One could relax it into a bilinear programming problem (by introducing many auxiliary variables) and solve the relaxation via branch and bound [20]. Instead we solve the inference approximately with two solvers, one is *alternating search*, and the other is *belief propagation* based on a novel factor graph representation.

3.6.1. Alternating search

This algorithm iteratively optimizes over the labelling space and the interaction structure space in turn. We initialize $\mathbf{e}$ with fully connected structures. During each iteration, we first make the graph structured $\mathbf{e}$ fixed, denoted by $\hat{\mathbf{e}}$, and solve the remaining problem:

$$\min_{\mathbf{a}} \sum_{i\in V}\sum_{s\in\mathbb{A}} \theta_{i;s}\mathbb{1}_s(a_i) + \sum_{i<j,\hat{e}_{ij}=1}\sum_{(s,t)\in\mathbb{A}^2} \theta_{i,j;s,t}\mathbb{1}_{s,t}(a_i, a_j). \tag{23}$$

The above is a CRF inference problem (with known graphs) and we solve it using the tree-reweighted message passing routine provided by the OpenGM package [31]. Denoting the current solution of $\mathbf{a}$ by $\hat{\mathbf{a}}$, and plugging it into (13), we get

$$\min_{\mathbf{e}} \sum_{i<j}\sum_{z\in\{0,1\}} \theta_{i,j;z}\mathbb{1}_z(e_{ij}) + e_{ij}\theta_{i,j;\hat{a}_i,\hat{a}_j} + \theta_0 e_{ij}. \tag{24}$$

Rearranging the terms within the objective of (24), we get the following equivalent form:

$$\min_{\mathbf{e}} \sum_{i<j} \left[\theta_{i,j;1} - \theta_{i,j;0} + \theta_{i,j;\hat{a}_i,\hat{a}_j} + \theta_0\right] e_{ij}. \tag{25}$$

Note for each (i, j), the value of e_{ij} does not affect the optimization of all rest e variables. Hence the optimal value of e_{ij} only depends on the sign of its coefficient. Since we are minimizing over e variables, the optimal value of e_{ij} is 1 (0) if the output of $\theta_{i,j;1} - \theta_{i,j;0} + \theta_{i,j;\hat{a}_i,\hat{a}_j} + \theta_0$ is negative (non-negative). Our alternating search procedure is summarized by Algorithm 1.

Algorithm 1: The alternating search algorithm.

Input: $\theta_{i,s}$, $\theta_{j,k;s,t}$, $\theta_{u,v;z}$, $\forall i, j, k, u, v \in V$, $s, t \in \mathbb{A}$, T.
Output: Action estimations and HIC estimations.
Initialization: Let $e_{ij}^0 = 1, \forall i, j \in V, t = 0$.
while $t < T$ **do**
  for $c \in C$ **do**
    Solve the optimization (23) with OpenGM [31] with the current HIC solution $\mathbf{e}^t$. Let $\mathbf{a}^t$ denote the solution.
    for $\forall i < j,\ i, j \in V$ **do**
      $\hat{e}_{ij}^{t+1} = 0$.
      Set $\hat{e}_{ij}^{t+1} = 1$ if $\theta_{i,j;1} - \theta_{i,j;0} + \theta_{i,j;\hat{a}_i^t,\hat{a}_j^t} + \theta_0 < 0$.
    end
  end
  $t = t + 1$.
end
Return $\hat{\mathbf{a}}^{T-1}$, $\hat{\mathbf{e}}^{T-1}$.

3.6.2. Belief propagation with third-order factor graph representation

It turns out that the inference problem (13) can be transferred into a Markov random field with higher-order potentials. Specifically, we define $\phi_i(a_i), \phi_{ij}(e_{ij}), \psi_{c_{ij}}(a_i, a_j, e_{ij})$ as:

$$\phi_i(a_i) = \theta_{i;a_i} \tag{26}$$

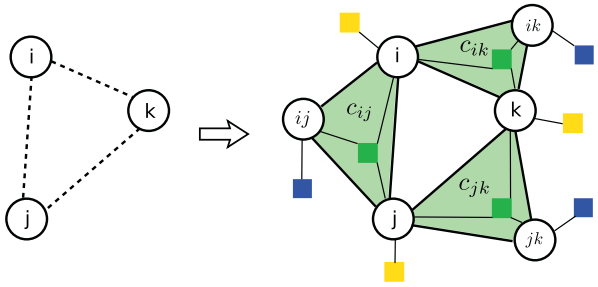


Fig. 2. Transfer the inference problem with unknown interactions to Markov random field with known interactions. Here factors are categorized into three types including action factors (yellow squares), interaction factors (blue squares) and action-interaction factors (green squares). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

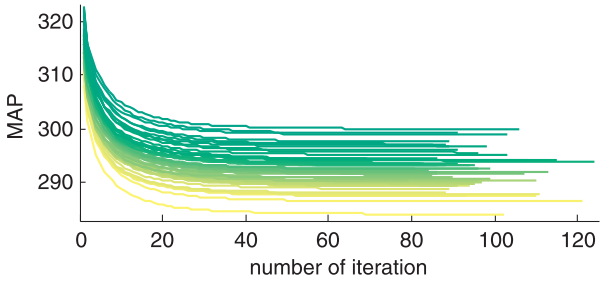


Fig. 3. Implementing Algorithm 2 on 50 inference problems by creating random potentials with a number of 20 nodes ($|V| = 20$). For these examples, the algorithm always converges in 125 iterations.

$$\phi_{ij}(e_{ij}) = \theta_{ij;e_{ij}} + \theta_o e_{ij} \tag{27}$$

$$\psi_{c_{ij}}(a_i, a_j, e_{ij}) = e_{ij}\theta_{i,j;a_i,a_j}. \tag{28}$$

Then one can reformulate (13) into:

$$\min_{\mathbf{a},\mathbf{e}} \sum_{i\in V}\phi_i(a_i) + \sum_{j>i}\phi_{ij}(e_{ij}) + \psi_{c_{ij}}(a_i, a_j, e_{ij})$$
$$\text{s.t. } a_i \in \mathbb{A},\ e_{ij} \in \{0,1\} \forall i < j. \tag{29}$$

This process is illustrated by Fig. 2, where the original inference problem with unknown interactions is equivalently transferred into a factor graph representation with known structure, by introducing a third-order potential function $\psi_{c_{ij}}(a_i, a_j, e_{ij})$. Specifically, this potential considers the correlation between the action categories of two individuals and their HIC (that is, the existence of interaction between i and j). The semantics of this potential is that if there exist any interaction between two individuals (*i.e.* $e_{ij} = 1$), then the compatibility between action labels (a_i and a_j) is evaluated. Otherwise ($e_{ij} = 0$), we do not need to consider their compatibility since there is no interaction between the associated people. With such reformulation, we are able to derive a different iterative algorithm from the *alternating search* presented in Section 3.6.1.

For each cluster c_{ij} (note c_{ij} and c_{ji} denote the same cluster), we introduce $\lambda_{c_{ij}\to i}(a_i)$, $\lambda_{c_{ij}\to j}(a_j)$ and $\lambda_{c_{ij}\to ij}(e_{ij})$ as Lagrangian multipliers. Let $N(i)$ denote all nodes neighboring to node i except for the interaction nodes in the factor graph in Fig. 2. By applying dual decomposition [32] to the LP-relaxation of problem (29), we get a convex dual problem of the relaxed problem shown below:

$$\min_{\lambda} \sum_{i\in V}\max_{a_i}\left[\phi_i(a_i) + \sum_{j\in N(i)}\lambda_{c_{ij}\to i}(a_i)\right] + \sum_{j>i}\max_{e_{ij}}\left[\phi_{ij}(e_{ij}) + \lambda_{c_{ij}\to ij}(e_{ij})\right]$$
$$+ \sum_{j>i}\max_{a_i,a_j,e_{ij}}\left[\psi_{ij}(a_i, a_j, e_{ij}) - \lambda_{c_{ij}\to i}(a_i) - \lambda_{c_{ij}\to j}(a_j) - \lambda_{c_{ij}\to ij}(e_{ij})\right]. \tag{30}$$

One can check that within (30), the objective is convex and the feasible set is a convex set, hence the optimization problem (30) is a convex optimization problem. Instead of solving the origin optimization (29), we select to solve this dual optimization by deriving a message-passing alike algorithm using the so-called coordinate descent strategy proposed by Globerson and Jaakkola [33]. Specifically, picking a particular cluster c_{ij}, and letting $\lambda_{c_{ij}\to i}(a_i)$, $\lambda_{c_{ij}\to j}(a_j)$ and $\lambda_{c_{ij}\to ij}(e_{ij})$ be flexible with all rest variables fixed, we get a sub-problem as follows:

$$\max_{a_i}\left[\phi_i(a_i) + \sum_{j'\in N(i)}\lambda_{c_{ij'}\to i}(a_i)\right] + \max_{a_j}\left[\phi_j(a_j) + \sum_{i'\in N(j)}\lambda_{c_{i'j}\to j}(a_j)\right]$$
$$+ \max_{e_{ij}}\left[\phi_{ij}(e_{ij}) + \lambda_{c_{ij}\to ij}(e_{ij})\right] + \max_{a_i,a_j,e_{ij}}\left[\psi_{ij}(a_i, a_j, e_{ij}) - \lambda_{c_{ij}\to i}(a_i)\right.$$
$$\left. - \lambda_{c_{ij}\to j}(a_j) - \lambda_{c_{ij}\to ij}(e_{ij})\right]. \tag{31}$$

It turns out that problem (31) has a closed-form solution. For each triplet (a_i, a_j, e_{ij}), we denote $\beta_{ij}(a_i, a_j, e_{ij})$ by:

$$\beta_{ij}(a_i, a_j, e_{ij}) = \psi_{ij}(a_i, a_j, e_{ij}) + \phi_{ij}(e_{ij}) + \phi_i(a_i) + \phi_j(a_j)$$
$$+ \sum_{j'\in N(i)\setminus\{j\}}\lambda_{c_{ij'}\to i}(a_i) + \sum_{i'\in N(j)\setminus\{i\}}\lambda_{c_{i'j}\to j}(a_j). \tag{32}$$

Then the closed-form solution of (31) is

$$\lambda^*_{c_{ij}\to i}(a_i) = -\phi_i(a_i) - \sum_{j'\in N(i)\setminus\{j\}}\lambda_{c_{ij'}\to i}(a_i) + \frac{1}{3}\max_{a_j,e_{ij}}\beta_{ij}(a_i, a_j, e_{ij}), \tag{33a}$$

$$\lambda^*_{c_{ij}\to j}(a_j) = -\phi_j(a_j) - \sum_{i'\in N(j)\setminus\{i\}}\lambda_{c_{i'j}\to j}(a_j) + \frac{1}{3}\max_{a_i,e_{ij}}\beta_{ij}(a_i, a_j, e_{ij}), \tag{33b}$$

$$\lambda^*_{c_{ij}\to ij}(e_{ij}) = -\phi_{ij}(e_{ij}) + \frac{1}{3}\max_{a_i,a_j}\beta_{ij}(a_i, a_j, e_{ij}). \tag{33c}$$

Here $\lambda_{c_{ij}\to j}(a_j)$ denotes the cluster c_{ij}'s belief when person j's action is a_j, and $\lambda_{c_{ij}\to ij}(e_{ij})$ denotes the same cluster's belief when person i and person j have interaction with each other. Taking the results in Eqs. (31)–(33), we solve the dual problem in an iterative manner shown by Algorithm 2. Since the dual problem is a convex optimization problem, and Algorithm 2 decreases the dual objective at each iteration, the algorithm guarantees to converge to a fixed point of the dual problem, see Fig. 3 for an empirical study of the optimization process of the algorithm. However, as pointed out by Globerson and Jaakkola [33], in general the fixed point fails to admit dual optimum (except for cases that all functions in (33) have unique maxima) as the coordinate-descent routine often gets stuck at local solutions.

4. Experiment and result

We first evaluate the inference algorithms with respect to accuracy and time costs. Then we investigate the proposed joint learning approach on human activity understanding dataset and compare it against the-state-of-arts if available.

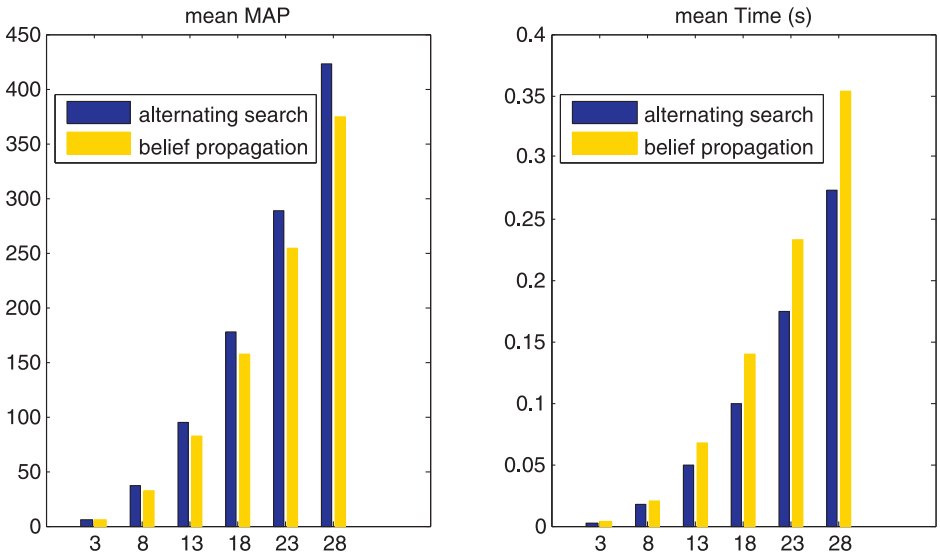


Fig. 4. Compare inference algorithms on synthetic data. Left diagram shows mean MAP values of 1000 samples, and right diagram gives mean time-costs, as the number of nodes increases from 3 to 28.

Algorithm 2: The belief propagation algorithm.

Input: Graph $\mathcal{G} = \left(\{V, V^+\}, \mathcal{C}\right)$, $\phi_i(a_i)\ \forall i \in V$, $\phi_{jk}(e_{jk})\ \forall j, k \in V^+$, $\psi_{c_{ij}}(a_i, a_j, e_{ij})\ \forall i, j \in V, i < j$.
Output: The action estimations $\mathbf{a}^*$ and HIC estimations $\mathbf{e}^*$.

Initialization:
Set $\lambda_{c_{ij}\to i}(a_i) = 0$, $\lambda_{c_{ij}\to j}(a_j) = 0$, $\lambda_{c_{ij}\to ij}(e_{ij}) = 0$ $\forall ij, a_i, a_j, e_{ij}$.
while *not convergent* **do**
  for $c \in \mathcal{C}$ **do**
    Update $\lambda_{c_{ij}\to i}(a_i), \lambda_{c_{ij}\to j}(a_j), \lambda_{c_{ij}\to ij}(e_{ij})$ via Eq. (33).
  end
end
Calculate $b_u(a_u) = \phi_u(a_u) + \sum_{\mathcal{C}; u\in\mathcal{C}} \lambda_{c\to u}(a_u),\ \forall u \in V$.
Compute $b_{ij}(e_{ij}) = \phi_{ij}(e_{ij}) + \lambda_{c_{ij}\to ij}(e_{ij})\ \forall (i, j)$.
Decoding: $a^* = \text{argmax}_a b(a)$, $e^* = \text{argmax}_e b(e)$.

4.1. Inference evaluation

We evaluate the proposed inference algorithms on 1,000 synthetic inference problems when the numbers of nodes ($|V|$) are set to 3, 8, 13, 18, 23, 28 respectively. For a fair comparison, both algorithms stop in 10 iterations.

Fig. 4 shows the results, from which we can see that *alternating search* outperforms *belief propagation* in terms of both objective value and time-consuming. The reason might be that *belief propagation* solves a relaxed optimization rather than the original problem (29). Hence the decoded solution in Algorithm 2 does not necessarily admit the optimal solution to the original optimization. In other words, the relaxation (*i.e.* the primal problem of (30)) is not tight enough to give a premium approximation to the original problem (29). One could argue that the relaxation can be tightened using the cluster pursuit technique in [34]. However, this will introduce additional computational cost on finding auxiliary clusters, and we simply get rid of this at present.

4.2. Activity understanding evaluation

We evaluate our approach on three commonly used dataset for human interaction understanding, which include UT, TVHI and BIT.

1. TVHI is a collection of 300 videos from 23 TV shows [35]. It contains 5 action classes: handshake (HS), highfive (HF), hug (HG), kiss (KS) and no-action (NO). For each of the first four classes, 50 short videos are collected, while for the no-interaction class, 100 clips are collected. The dataset is challenging because many videos contain complex background and cluttered foreground. Besides, the dataset is severely biased in terms of numbers of examples of different action classes. We split the dataset into two parts in the way suggested by Patron-Perez et al. [35], and then use them to train models and to do testing in turn, which ensures the evaluation is performed on the whole dataset.
2. UT [36] contains 120 short videos of 6 action classes: handshake, hug, kick, punch and push. UT includes three asymmetrical actions: kick, punch and push. As done in [13], we add a complementary class for each asymmetrical class. As a result, we have 9 action classes including no-action (NO), handshake (HS), hug (HG), kick (KK), be-kicked (BKK), push (PS), be-pushed (BPS), punch (PC), be-punched (BPC). As suggested by Wang et al. [13] the dataset is split into 2 subsets used for training and testing respectively. Again, the power of training interactions jointly with actions can be verified by comparing results of the proposed approach and results of sparse and dense CRFs.
3. BIT [37] contains 400 video clips and 8 action classes: bow, boxing, handshake, highfive, hug, kick, pat, and push. As done for UT, we add complementary classes for asymmetrical actions, and we have 14 action classes in total including no-action (NO), bend (BD), be-bent (BBD), box (BX), be-boxed (BBX), handshake (HS), highfive (HF), hug (HG), kick (KK), be-kicked (BKK), pat (PT), be-patted (BPT), push (PS) and be-pushed (BPS). We use half of the dataset for training, and the other half for testing.

We denote our approaches by *joint+AS* and *joint+BP*, which use the same joint training framework (Section 3) but solve the relevant inference problems via alternating search (AS) and belief propagation (BP) respectively. We compare the two methods with baselines and the state-of-the-arts. Baselines are:

1. *HoG+HoF*: the linear SVM model trained with the HoG+HoF descriptor used by Wang et al. [13];
2. *Spatial Net*: the spatial net proposed by Simonyan and Zisserman [5] for action recognition;

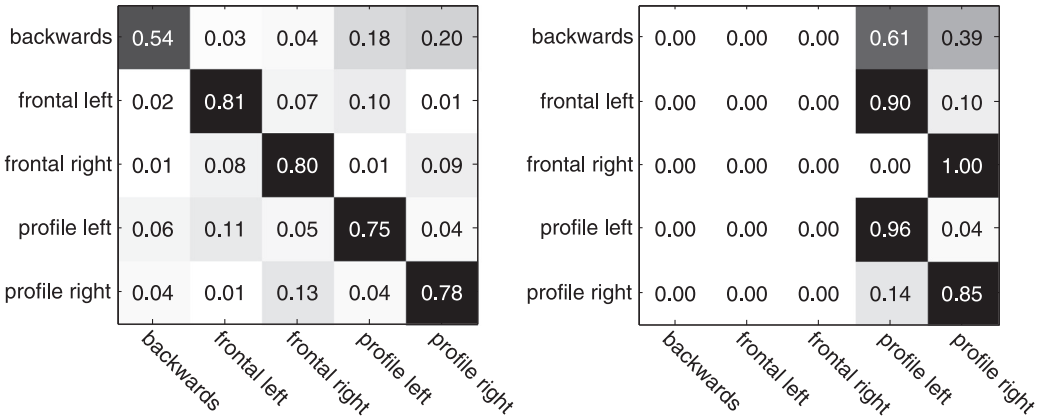


Fig. 5. Confusion matrix for head orientation recognition on TVHI (left) and UT (right). *profile-left* and *profile-right* classes dominate UT data, hence accuracies across different classes are significantly biased.

3. *Combined+SVM* (cmb+SVM): we concatenate the outputs of FC6 of the Spatial Net, the HoG and HoF descriptors to form the action descriptors, which are then used to train linear SVM classifiers;
4. *dense CRF+Combined* (dense CRF+Cmb): the CRF model (with fixed complete graphs) includes unary and pairwise energies, where the unary energy is computed with the *Combined* descriptors.

The state-of-the-arts include:

1. *sparse CRF*: the CRF model proposed recently in [13]. Both the interactions and actions are learned separately from data, which is different from the joint learning approach proposed in this paper.
2. *sparse CRF+Combined* (sparse CRF+Cmb): this models shares learned interactions with *sparse CRF*. Different from *sparse CRF*, it uses the combined descriptors and the head orientation to construct the energy function.
3. *latent CRF+Combined* (latent CRF+Cmb): the approach proposed in [19], which treats HIC as latent variables and estimates actions and interactions simultaneously, see Section 2.2 for more details. This method share the same unary and pairwise energy temrs with our approach for fair comparison.

4.2.1. Implementation details

Training CNNs is more challenging for action recognition due to facts that recognizing actions is more complicate than recognizing objects, the amount of available dataset is typically limited, and training samples over different classes are imbalanced. To ease the training of deep models, we pre-train CNN models on ImageNet. Before training we augment ImageNet by three operations: 1) all images are rescaled to 256×256; 2) we randomly crop 224×224 sub-images from original images; 3) then the generated images undergo horizontal flipping. The network weights are learned using the mini-batch stochastic gradient descent with momentum setting to 0.9. The learning rate is initially set to 0.01, then multiply by 0.1 after every 100k iterations. The training stops after 450k iterations.

Before fine-tuning on datasets for action recognition, we augment the data by scaling each image by three different factors $\{0.9, 1.0, 1.1\}$, afterwards rotating the scaled data by $\{-5, 0, 5\}$ degrees. We then fine-tune the pre-trained model on the augmented data, with the momentum and the dropout ratio set to 0.9, 0.5 respectively. To fine-tune the pre-trained model on TVHI, the step sizes are set to 8000 and 5000, the max-numbers-of-iteration are 24k and 15k, for action and interaction recognition respectively. To do fine-tuning on UT, the step sizes are 2000 and 1000, and the max-numbers-of-iteration are 10k and 5k. For BIT, the step sizes are 5000 and 3000, and the max-numbers-of-iteration are 15k and 10k.

In order to estimate human-head-orientations, we first use the CNN proposed in [29] to extract features from the areas of heads (we estimate head locations empirically within human-body bounding boxes). We then concatenate the extracted deep features with HoG descriptors to obtain the final descriptors, which are then taken to train orientation classifiers with linear SVM. The classification results on TVHI and UT are shown in Fig. 5. For BIT, we do not use the head orientation within all CRF models as heads are typically very small here and estimating their orientations is difficult.

Both HoG and HoF extractions on TVHI use the 8×8-grid setting and 5-orientational bins. HoG extraction on UT uses the 6×12-grid setting and 5-orientational bins. HoF extraction on UT uses the 6×12-grid setting as well but 9-orientational bins.

To train SVM classifiers, we use Liblinear toolbox [38]. All structured models are trained using the structured-SVM toolbox provided by Joachims [39]. All latent structured models are trained using the toolbox by Do and Artieres [24]. For loopy belief propagation, we use the OpenGM package [31].

4.2.2. Human body detection

Since Yolo V2 [40] achieves the state-of-the-art performance on object detection, and is very fast, we take it to detect human bodies in this paper (here we use the default threshold 0.24). Instead of using the detected bounding boxes directly, we further refine the detection by Yolo V2 with two additional steps: 1) we discard the bounding boxes which are smaller than a size of $\frac{1}{3} \times$ area of the largest bounding box within each frame; 2) then select 4 bounding boxes with largest confidences among all detections generated by the first step. Empirically, this post-processing step filters irrelevant human bodies for the recognition task and keeps informative detections as that shown in Fig. 7.

In order to quantitatively evaluate the performance on detection, we compute the precision/recall (PR) curve for each dataset and show the result in Fig. 6. It can be seen that detections using Yolo V2 and our refining process are superior though not perfect. Also, the performance on BIT and UT is obviously better than TVHI. This is because occlusion in TVHI is much more serious than BIT and UT, meanwhile background in TVHI is more complicated.

4.2.3. Action and interaction recognition on TVHI

Results are shown in Table 1. It can be seen that CNN descriptors outperform handcrafted features by a large margin with respect to both action and interaction recognition. Further, the

Table 1
Action recognition accuracy (%) on TVHI.

Method	No-action	Handshake	Highfive	Hug	Kiss	No-interacting	Interacting	Overall	Mean
HoG+HoF	52.4	26.5	25.7	20.0	36.1	–	–	44.5	32.1
Spatial Net [5]	54.7	38.4	35.4	54.8	58.6	–	–	53.6	48.4
Cmb+SVM	59.9	41.7	46.0	56.2	66.7	–	–	58.5	54.1
dense CRF+Cmb	70.4	32.0	32.0	55.5	**70.0**	0	**100**	47.3	51.4
sparse CRF [13]	82.1	16.0	14.0	30.1	42.7	73.5	52.5	67.1	43.6
latent CRF [19]+Cmb	**86.0**	28.3	22.1	45.1	43.1	78.2	48.7	**71.6**	50.2
sparse CRF [13]+Cmb	59.9	**46.9**	43.8	58.3	63.8	84.7	62.8	67.5	60.0
joint + AS (ours)	57.9	46.0	**46.9**	**63.1**	67.9	**86.2**	67.4	68.1	**62.2**
joint + BP (ours)	57.3	46.3	46.4	62.5	68.4	85.6	67.8	67.7	62.0

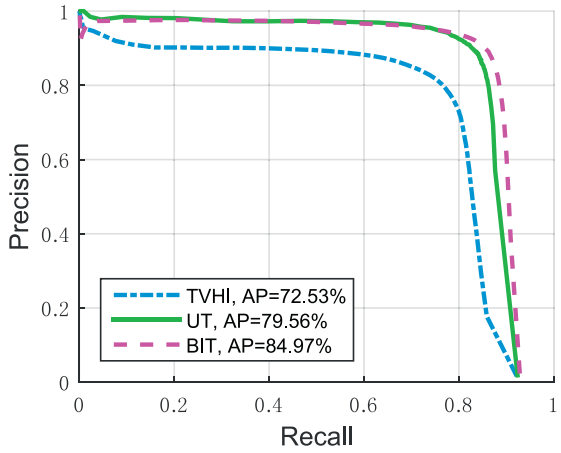


Fig. 6. Detection results on TVHI, UT and BIT. Both PR curve and average precision (AP) are given.

concatenation of CNN descriptor and handcrafted feature outperforms each of them on action recognition, as shown in Table 1. Overall the proposed joint learning approach (joint with *alternating search* and *joint with belief propagation*) outperforms all state-of-the-arts except for *latent CRF+Combined*.

Conversely, with respect to the mean accuracy, our approach outperforms *latent CRF+Combined* by a large margin (12%). This is because that the dataset is dominated by the no-action class (accounts around 69% for all examples), on which *latent CRF+Combined* performs much better than our approach, whereas our approach outperforms *latent CRF+Combined* significantly on all rest classes. In conclusion our approach admits much unbaised results across different classes compared against latent CRF. By comparing the results of our approach (68.1% for overall, 62.2% for mean) and results by *dense CRF* (47.3% for overall, 51.4% for mean) and *sparse CRF* (67.5% for overall, 60.0% for mean), we can conclude that the joint training of interactions and actions performs better (on action recognition) than using static hand-crafted interaction-configurations. Besides, it is easy to find that our approach achieves the best results with respect to interaction recognition as well (see *no-interacting* and *interacting* columns). With the recognition results on both action and interaction, our approach is able to deliver a good understanding of human activities in videos.

4.2.4. Action and interaction recognition on UT

For UT we use the same evaluation protocol as TVHI. Results are shown in Table 2. The proposed joint learning approach outperforms the second best method *sparse CRF+Combined* by(0.5%) and (1.3%) with respect to the overall and mean accuracy respectively.

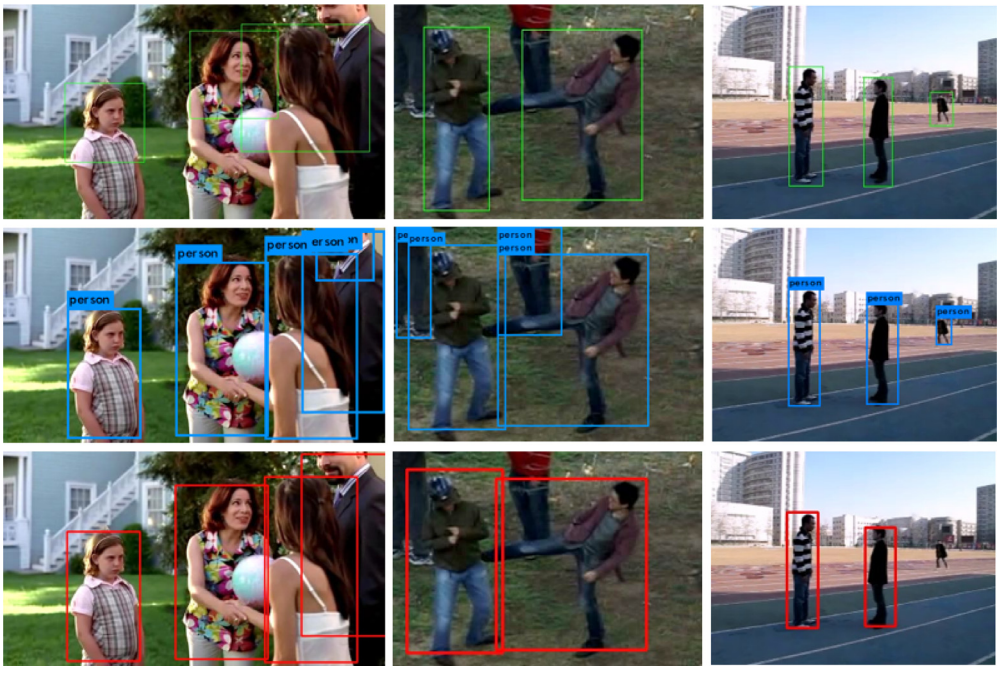


Fig. 7. Human detection and selection results on TVHI (left), UT (middle) and BIT (right). The first row is ground-truth; the second row is the detected results using Yolov2; the last row is the selected bounding boxes with the approach in Section 4.2.2.

Table 2
Action recognition accuracy (%) on UT. Here NO, HS, HG, KK, BKK, PS, BPS, PC, BPC, NINT, INT represent no-action, handshake, hug, kick, be-kicked, push, be-pushed, punch, be-punched, non-interacting and interacting respectively.

Method	NO	HS	KK	HG	PC	PS	BKK	BPC	BPS	NINT	INT	Overall	Mean
HoG+HoF	81.9	93.6	70.0	97.4	32.1	51.7	39.1	35.9	56.6	–	–	77.0	62.1
Spatial Net [5]	54.8	77.1	67.4	97.8	18.0	41.4	13.0	28.2	51.7	–	–	63.9	50.0
Combined+SVM	84.1	97.6	82.6	98.1	42.3	41.3	59.8	61.5	84.8	–	–	83.0	72.5
dense CRF+Cmb	90.5	99.3	90.2	98.1	43.6	62.1	72.8	57.7	84.8	0	**100**	84.1	72.6
sparse CRF [13]	**92.9**	99.3	89.1	**100**	35.9	60.7	62.0	32.1	60.7	99.0	99.1	89.3	75.5
latent CRF [19]+Cmb	66.7	84.8	85.9	84.9	**53.9**	46.9	39.1	**66.7**	**92.4**	**100**	39.7	67.2	69.2
sparse CRF [13]+Cmb	84.1	99.3	90.2	98.1	43.6	61.4	71.7	57.7	84.1	**100**	99.9	90.5	80.9
joint + AS (ours)	84.1	**99.5**	**91.3**	98.8	51.3	62.8	**77.2**	59.0	80.7	**100**	99.9	**91.0**	**82.2**
joint + BP (ours)	84.1	**99.5**	**91.3**	98.8	44.9	**64.1**	**77.2**	59.0	83.5	**100**	100	**91.0**	82.0

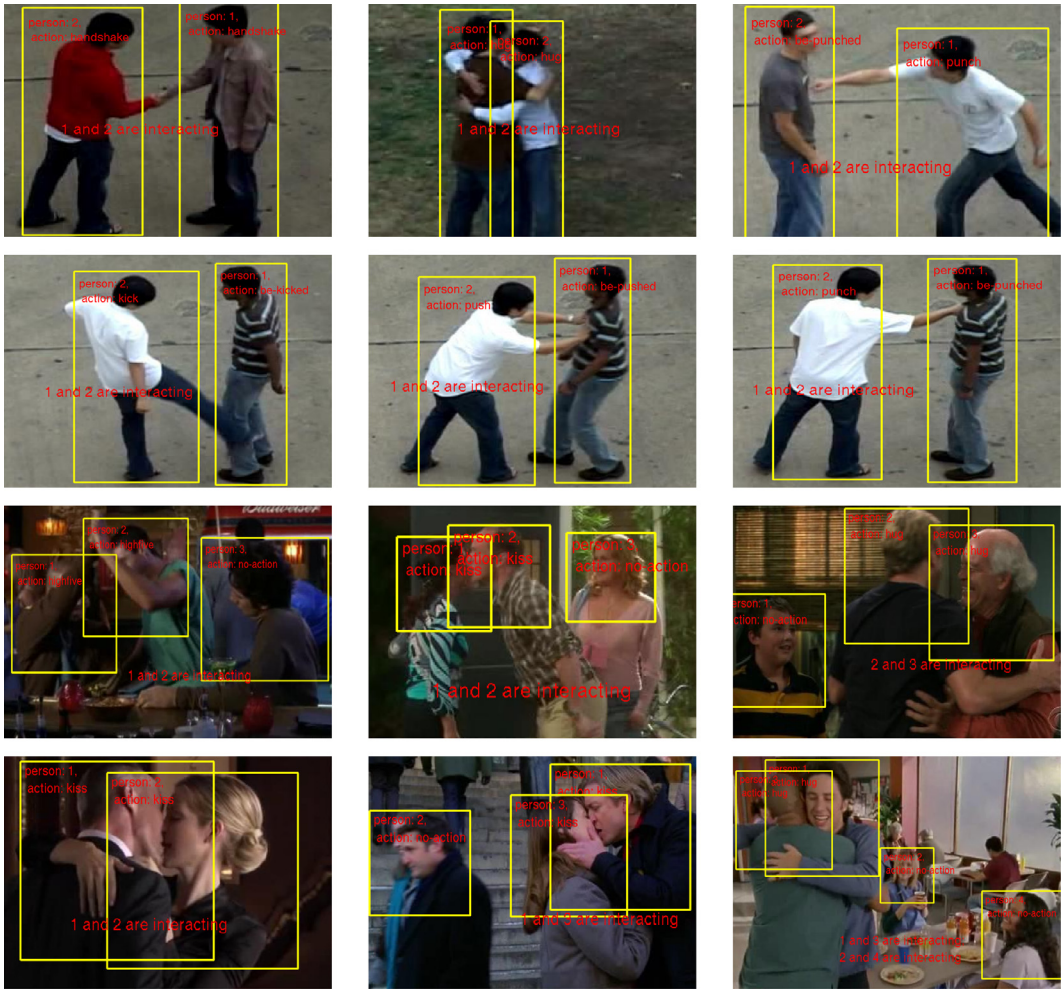


Fig. 8. Human interaction understanding using our recognition results. Yellow rectangles represent detected human bodies. The text superimposed onto each image tell what happens in that scene, which is generated by combining action and interaction recognition results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The results again justify that the proposed method is competitive compared with the state-of-the-arts.

4.2.5. Action and interaction recognition on BIT

For BIT again we use the same evaluation protocol as TVHI. Results are shown in Table 3. Again the proposed approach outperforms baseline methods significantly with respect to action classification, and is competitive compared with the state-of-the-arts.

Compared with a majority of existing approaches which only predict each video an action label, our framework enables a semantic understanding of what happens in images or videos, see Fig. 8. Specifically, we can take the action and interaction recognition results to describe the activity in each frame semantically (for instance, *person 1* is punching *person 2*). For *dense CRF* or *sparse CRF*, the HICs are determined heuristically, and generating such semantic understanding of human activities can be inaccurate with such CRF representations.

4.2.6. The impact of inference

We now compare the proposed inference techniques on real data, namely *alternating search* and *Belief propagation*. From results in Tables 1–3, we can see that *alternating search* performs slightly better than *Belief propagation* in general. Indeed, the number of nodes in real data is typically less than 4, and the

Table 3
Action recognition accuracy (%) on BIT. Here NO, BD, BBD, BX, BBX, HS, HF, HG, KK, BKK, PT, BPT, PS, BPS, NINT, INT represent no-action, bend, be-bent, box, be-boxed, handshake, highfive, hug, kick, be-kicked, pat, be-patted, push, be-pushed, non-interacting and interacting respectively.

Method	NO	BD	BBD	BX	BBX	HS	HF	HG	KK	BKK	PT	BPT	PS	BPS	NINT	INT	Overall	Mean
HoG+HoF	70.2	78.2	16.2	**24.6**	21.8	39.8	60.6	39.3	77.6	17.0	56.5	20.4	22.8	23.7	–	–	60.3	40.6
Spatial Net [5]	83.3	68.0	2.1	2.6	14.2	13.3	47.4	70.9	62.1	4.0	44.1	21.1	19.6	12.4	–	–	67.7	33.2
Combined+SVM	82.7	90.1	3.9	9.7	**25.9**	45.8	67.0	67.1	87.3	14.6	57.6	35.2	34.5	28.3	–	–	71.7	46.4
dense CRF+Cmb	87.9	90.5	37.9	10.9	22.0	47.0	72.7	62.3	88.6	24.3	68.8	56.6	36.7	34.2	0.0	**1.0**	42.2	52.5
sparse CRF [13]	87.6	91.2	30.7	10.4	21.7	50.1	72.1	63.0	88.8	23.7	68.0	56.2	36.8	34.8	13.9	99.5	48.9	53.0
latent CRF [19]+Cmb	**93.9**	92.7	0.0	6.9	8.9	1.3	45.8	16.8	**95.9**	23.7	57.2	2.2	32.5	10.0	**99.9**	14.5	80.9	37.6
sparse CRF [13]+Cmb	83.5	91.2	61.6	12.2	21.8	43.8	**73.8**	68.5	87.5	38.3	67.7	59.3	35.8	34.7	94.5	88.3	84.8	60.2
joint+AS (ours)	80.0	92.7	59.7	18.3	22.9	62.9	73.1	73.2	83.2	**47.4**	67.5	62.0	**38.3**	**37.4**	96.7	81.1	**85.0**	62.3
joint+BP (ours)	79.3	**93.9**	**61.8**	17.7	23.2	**64.5**	69.7	**76.9**	81.2	42.9	**72.1**	**66.7**	35.5	35.1	95.9	85.0	84.8	**62.6**

Table 4
Action recognition accuracy with detection (%) on TVHI, UT and BIT.

Dataset	Method	Overall	Mean
TVHI	sparse CRF + Cmb	**65.8**	55.5
	joint + AS	64.6	**56.6**
	joint + BP	64.4	56.0
UT	sparse CRF+Cmb	89.6	71.4
	joint + AS	**90.3**	73.7
	joint + BP	90.2	**74.8**
BIT	sparse CRF+Cmb	81.1	54.9
	joint + AS	82.8	58.8
	joint + BP	**83.0**	**60.0**

performance of two algorithms on problems of this scale is similar. This phenomenon coincides with the evaluation on synthetic data, see Fig. 4-left when the number of nodes is 3. We believe that as the scale of the problem increases, the difference between two inference algorithms will become salient.

4.2.7. The impact of human-body detection

We now evaluate how human-body detection affects the recognition performance. We use the same model as before while replacing the annotated bounding boxes with detected ones by Yolov2. Table 4 gives the results, from which we conclude: 1) accurate detections are important to action recognition, as recognition with annotated bounding boxes performs much better than detected ones; 2) with detected human bodies, our joint training approach still outperforms the methods using heuristic interactions, *i.e. sparse CRF+combined*.

5. Conclusion

We have presented an approach to learn the configuration of human interactions and their action labels in a unique framework. Our formulation learns both actions and human interactions explicitly and simultaneously in a supervised manner, using both deep-neural-network representations and high-level contextual information. We used a max-margin-style training approach to train model parameters, and proposed two efficient and effective optimization algorithms (alternating search and belief propagation) to solve the relevant inference problem. We compared their performance and found that the simple alternating search algorithm outperforms the belief propagation-style algorithm on both synthetic and read data. Indeed, understanding human activities in videos can benefit from our joint learning approach, as that evidenced by comparison with baselines and the state-of-the-arts.

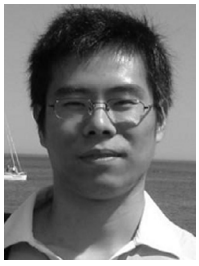
Acknowledgments

Zhenhua Wang was partially supported by National Natural Science Foundation of China (61802348) and Zhejiang Provincial Natural Science Fundation of China (LQ16F030007). Sheng Liu was partially supported by the Zhejiang Provincial Natural Science Fundation of China (LY15F020031). Shengyong Chen was supported in part by National Natural Science Foundation of China (U1509207, 61325019). Dongyan Guo was partially supported by the Zhejiang Provincial Natural Science Fundation of China (GB18041190041). Zhanpeng Shao was partially supported by National Natural Science Foundation of China (61603341).

References

[1] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, H.T. Shen, Unsupervised deep hashing with similarity-adaptive and discrete optimization, IEEE Trans. Pattern Anal. Mach. Intell. (2018), doi:10.1109/TPAMI.2018.2789887.
[2] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: multi-path refinement networks with identity mappings for high-resolution semantic segmentation, in: Proceedings of the CVPR (2017).
[3] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, Q. Dai, Supervised hash coding with deep neural network for environment perception of intelligent vehicles, IEEE Trans. Intel. Transp. Syst. PP (99) (2017) 1–12.
[4] R. Yao, G. Lin, Q. Shi, D. Ranasinghe, Efficient dense labelling of human activity sequences from wearables using fully convolutional networks Pattern Recognition 78 (2018) 252–266.
[5] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the NIPS, 2014.
[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. van Gool, Temporal segment networks: towards good practices for deep action recognition recognition, in: Proceedings of the ECCV, 2016.
[7] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.
[8] E. Park, X. Han, T. Berg, A. Berg, Combining multiple sources of knowledge in deep cnns for action recognition, in: Proceedings of the Winter Conference on Applications of Computer Vision (WACV), 2016.
[9] T. Du, L. Bourdev, R. Fergus, L. Torresani, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the ICCV, 2015.
[10] A. Kar, N. Rai, K. Sikka, G. Sharma, Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos, in: Proceedings of the CVPR, 2017.
[11] M. Amer, P. Lei, S. Todorovic, Hirf: Hierarchical random field for collective activity recognition in videos, in: Proceedings of the ECCV, 2014.
[12] T. Kaneko, M. Shimosaka, S. Odashima, R. Fukui, T. Sato, Consistent collective activity recognition with fully connected crfs, in: Proceedings of the ICPR, 2013.
[13] Z. Wang, S. Liu, J. Zhang, S. Chen, Q. Guan, A spatio-temporal CRF for human interaction understanding, IEEE Trans. Circuits Syst. Video Technol. 27 (8) (2017) 1647–1660.
[14] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, F. Wu, Efficient parallel framework for HEVC motion estimation on many-core processors, IEEE Trans. Circuits Syst. Video Technol. 24 (12) (2014) 2077–2089.
[15] C. Yan, Y. Zhang, J. Xu, F. Dai, L. Li, Q. Dai, F. Wu, A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors, IEEE Signal Process. Lett. 21 (5) (2014) 573–576.
[16] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the CVPR, 2008.
[17] W. Choi, K. Shahid, S. Savarese, Learning context for collective activity recognition, in: Proceedings of the CVPR, 2011.
[18] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Comput. Vis. Image Underst. 150 (C) (2016) 109–125.
[19] T. Lan, Y. Wang, W. Yang, S.N. Robinovitch, G. Mori, Discriminative latent models for recognizing contextual group activities, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1549–1562.
[20] Z. Wang, Q. Shi, C. Shen, A. van den Hengel, Bilinear programming for human activity recognition with unknown mrf graphs, in: Proceedings of the CVPR, 2013.
[21] Y. Wang, G. Mori, Max-margin hidden conditional random fields for human action recognition, in: Proceedings of the CVPR, 2009.

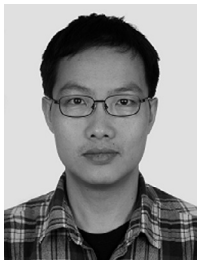
[22] C. Sminchisescu, A. Kanaujia, D. Metaxas, Conditional models for contextual human motion recognition, Comput. Vis. Image Underst. 104 (2) (2006) 210–220.
[23] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
[24] T.-M.-T. Do, T. Artieres, Large margin training for hidden markov models with partially observed states, in: Proceedings of the ICML, 2009.
[25] J. Jin, Z. Wang, S. Liu, J. Zhang, S. Chen, Q. Guan, Joint label-interaction learning for human action recognition, in: Proceedings of the ICIP, 2017.
[26] A. Patron-Perez, M. Marszalek, I. Reid, A. Zisserman, Structured learning of human interactions in tv shows, IEEE Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2441–2453.
[27] C.-N.J. Yu, T. Joachims, Learning structural svms with latent variables, in: Proceedings of the ICML, 2009.
[28] H. Wang, C. Schmid, Action recognition with improved trajactories, in: Proceedings of the ECCV, 2013.
[29] B. Ahn, J. Park, I. Kweon, Real-time head orientation from a monocular camera using deep neural network, in: Proceedings of the ACCV, 2015.
[30] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, J. Mach. Learn. Res. 6 (2) (2006) 1453–1484.
[31] B. Andres, T. Beier, J.H. Kappes, Opengm, 2015, (http://hciweb2.iwr.uni-heidelberg.de/opengm/index.php).
[32] N. Komodakis, N. Paragios, G. Tziritas, Mrf energy minimization and beyond via dual decomposition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 531–552.
[33] A. Globerson, T.S. Jaakkola, Fixing max-product: Convergent message passing algorithms for map lp-relaxations, in: Proceedings of the NIPS, 2007.
[34] Z. Zhang, Q. Shi, Y. Zhang, C. Shen, A.V.D. Hengel, Constraint reduction using marginal polytope diagrams for map lp relaxations, arXiv preprint: 1312.4637 (2013).
[35] A. Patron-Perez, M. Marszalek, A. Zisserman, I. Reid, High five: recognising human interactions in tv shows, in: Proceedings of the British Machine Vision Conference, 2010.
[36] M.S. Ryoo, J.K. Aggarwal, UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA), 2010, (http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html).
[37] Y. Kong, Y. Jia, Y. Fu, Learning human interaction by interactive phrases, in: Proceedings of the European Conference on Computer Vision, 7572, 2012, pp. 300–313.
[38] C. Lin, Liblinear – a library for large linear classification, 2015, (https://www.csie.ntu.edu.tw/~cjlin/liblinear/).
[39] T. Joachims, Support vector machine for complex outputs, 2008, (https://www.cs.cornell.edu/people/tj/svm_light/svm_struct.html).
[40] J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, in: Proceedings of the CVPR, 2017.



Zhenhua Wang is a lecturer in the College of Computer Science, Zhejiang University of Technology. He received a Ph.D. in Computer Vision in 2014 from The University of Adelaide. His research interests include computer vision, statistical learning and pattern recognition.

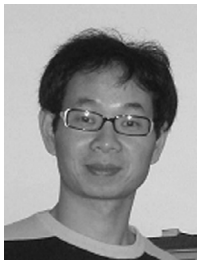
Jiali Jin received a Master's degree in Computer Science in 2018 from Zhejiang University of Technology. Her research interests include computer vision and machine learning.

Tong Liu received a Bachelor's degree in Computer Science in 2014 from Zhejiang University of Technology. His research interests include computer vision and artificial intelligence.



Sheng Liu received his Ph.D. from Zhejiang University in 2004. He is currently an associate professor in the Department of Computer Science, Zhejiang University of Technology. His research interests are video-based 3d reconstruction and Non-rigid Object Tracking.

Jianhua Zhang received his Ph.D. degree at the University of Hamburg in 2012. Now he works with College of Computer Science, Zhejiang University of Technology. His research interests include visual learning for autonomous robot, category discovery, object detection, image segmentation, medical image analysis. He is a member of the IEEE.



Shengyong Chen received a Ph.D. degree in computer vision from City University of Hong Kong, Hong Kong in 2003. He joined Zhejiang University of Technology, China, in Feb. 2004, where he is currently a Professor in the Department of Computer Science. His research interests include computer vision, robotics, and image analysis. He is a senior member of IEEE. He received the National Outstanding Youth Foundation Award of China in 2013.

Zhen Zhang is now a postdoctoral research fellow in the Department of Computer Science. Previously, he received a Ph.D. in Computer Science from Northwestern Polytechnical University. During Nov. 2012 to Dec. 2014, he was a visiting student at Australian Centre for Visual Technologies, the University of Adelaide. His research interests include machine learning and computer vision.

Dongyan Guo received the Bachelor's degree in application mathematics and the Ph.D. degree in pattern recognition and intelligent system from Nanjing University of Science and Technology, China, in 2008 and 2015, respectively. Since 2015, He has been a faculty member in the College of Computer Science, Zhejiang University of Technology. His research interests include computer vision and machine learning.

Zhanpeng Shao obtained a Ph.D. degree in computer vision from City University of Hong Kong, Hong Kong, in 2015. From 2015 to 2016, he was a senior research associate with the City University of Hong Kong Shenzhen Research Institute. He is currently an associate professor with the College of Computer Science and Technology. His research interests include computer vision, pattern recognition and machine learning.