JOINT LABEL-INTERACTION LEARNING FOR HUMAN ACTION RECOGNITION

Jiali Jin, Zhenhua Wang, Sheng Liu, Jianhua Zhang, Shengyong Chen and Qiu Guan*

Zhejiang University of Technology
288 Liuhe Rd. Hangzhou, P. R. China, 310023

ABSTRACT

Human interactions and their action categories preserve strong correlations, and the identification of the interaction configuration is of significant importance to improve the action recognition result. However, interactions are typically estimated using heuristics or treated as latent variables. The former usually produces incorrect interaction configuration while the latter introduces challenging training problem. Hence we propose a framework to jointly learn interactions and actions by designing a potential function using both features learned via deep neural networks and human interaction context. We propose an iterative approach to solve the associated inference problem efficiently and approximately. Experimental results on real datasets demonstrate that the proposed approach outperforms baselines by a large margin, and is competitive compared with the state-of-the-arts.

Index Terms— Structured prediction, conditional random field, human action recognition, human interaction

1. INTRODUCTION

Recognizing human actions is a fundamental problem in computer vision, and is essential for many applications such as sports video analysis, surveillance systems and video retrieval. Recent progress on deep learning boosts the action recognition performance significantly [1, 2, 3, 4, 5]. However, these works are not suitable to data that contains multiple people with interactions: they predict each image a single action label, hence are not applicable the multiple people scenarios. With respect to the multiple people with interactions, the interacting relations among them provide critical contextual information for recognizing complex human activities like handshake, fighting and football game.

In order to improve the performance of action recognition, researchers exploited conditional random fields (CRFs) [6, 7, 8] to model human interacting context. The associated potential functions of CRFs typically contain unary and pairwise terms, where the unary potential tells how much a particular action label matches the observed information extracted from image, while the pairwise potential evaluates the consistency of the labelling of its associated variables.

* Corresponding author.

To represent interactions among people, most CRF models use predefined graph structures, which are typically determined using domain knowledge or heuristics, *e.g*. the distance between persons [6], which are not robust against data variations and the change of imaging conditions. Using such graphs typically leads to bad recognition performance [9]. A better strategy is to learn CRF graphs from data. However, the human interaction structures are typically heterogeneous, which renders the traditional graph learning approach (see Chapter 18 in [10]) invalid. Lan *et al*. propose to estimate interactions jointly with actions via latent structured SVM [9], where the interactions are treated as latent variables, resulting in a non-convex training problem. Wang *et al*. use the same latent strategy as [9] to estimate human interactions and actions, with the inference problem solved by a slow branch and bound algorithm [11]. Very recently they determine the graph structure by classifying *interactions* and *non-interactions* with linear SVM classifiers [12]. Then the classifications are taken to construct the graph structures in their proposed spatial-temporal CRFs.

Our task is to obtain a frame-wise prediction of each person and their interaction configuration. The closest method to ours is the structured learning approach [13], where interactions and actions are learned jointly in a supervised manner without using latent variables. 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 delivers superior recognition results. Moreover, to solve the related inference algorithm, they resort to exhaustive search, while we propose a new inference algorithm that can scale to large optimization problems.

Our contributions are 1) we propose a novel training framework that is able to learn interactions and actions jointly without using latent variables; 2) we propose an algorithm to solve the associated inference problem efficiently; 3) the proposed approach outperforms the baseline methods by a large margin, while is very competitive in comparison with the state-of-the-arts.

2. JOINT INTERACTION AND ACTION LEARNING

Let $G = (V, E)$ be a graph, with the node set V representing the actions of all persons and the edge set E representing their interaction configuration. For instance $e_{ij} \in E$ means that person i and person j have interaction with each other, while the absent of the edge e_{st} indicates person s and person t are not interacting. Let $\mathbf{I}$ denote an image. Let $a_i \in \mathbb{A}$ be the action label of person i, $\mathbf{a} = [a_i]_{i=1,\dots,n}$ be a vector including the action labels of n people.

Given a new input $\mathbf{I}$, our aim is to predict both the action labelling $\mathbf{a}$ and the interaction configuration G via solving

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

where

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

where $\mathbb{1}_s(a_i)$ is an indicator function which outputs 1 if $a_i = s$, and outputs 0 otherwise. The indicator function $\mathbb{1}_z(e_{ij})$ gives 1 if $e_{ij} = z$ (and 0 otherwise). Likewise, $\mathbb{1}_{s,t}(a_i, a_j)$ is another indicator function that outputs 1 if $a_i = s$ and $a_j = t$, and gives 0 otherwise.

2.1. Unary energy

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

$$\theta_{i;s} = -p_{i;s}\, \mathbf{w}_u^\top\, \mathbf{1}_u(s), \tag{3}$$

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 a-th 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 features learned with deep convolutional neural networks (CNNs) [1], the histogram of gradients (HOG) features and the histogram of optical flow (HOF) features extracted from human body areas. We extract warped optical flow using the method described in [14].

2.2. Pairwise energy

The pairwise energy is given by

$$\theta_{i,j;s,t} = -\,\mathbf{w}_c^\top\, \mathbf{1}_c(s, t; l_{i,j}; p_i, p_j). \tag{4}$$

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.

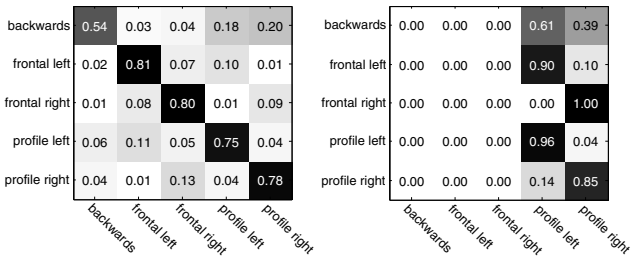

Fig. 1. Confusion matrix for head orientation recognition on TVHI (left) and UT (right). *profile-left* and *profile-right* classes dominate UT data, hence the result is significantly biased.

Here $l_{i,j} \in \mathbb{L}$ denotes the relative distance of person i to person j, which is

$$l_{i,j} = \lceil h_{i,j} / \bar{h} \rceil. \tag{5}$$

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

$$h_{i,j} = \frac{2 d_{i,j}}{r_i + r_j}. \tag{6}$$

In Equation (5), $\bar{h}$ denotes the mean h value of all interacting pairs (i, j) selected from training samples.

As typically done in action recognition, 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 one person, we train a linear large-margin classifier using both hand-crafted feature and the feature learned with a ConvNet [15]. The classification results are shown in Figure 1.

2.3. Interaction energy

The energy provides the negative confidence with respect to the existence of interaction among people. The definition is

$$\theta_{i,j;z} = -\{p_{i,j;z}\, \mathbf{w}_\tau^\top\, \mathbf{1}_\tau(z) + \mathbf{w}_r^\top\, \mathbf{1}_r(l_{i,j}; p_i, p_j; z)\}. \tag{7}$$

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 a SVM classifier using features extracted from areas that enclose bouning boxes of i and j. The feature representation here is similar to the features used to compute $p_{i;s}$ within the unary energy. $\mathbf{1}_\tau(z) \in \{0, 1\}^2$ is an indicator vector which takes 1 at its z-th position, and takes 0 elsewhere.

The parameter $\mathbf{w}_r$ evaluates the compatibility among the interaction configuration (encoded by z), the relative distance $l_{i,j}$ and the head orientations p_i, p_j. The purpose here it to leverage these contextual cues to improve the discrimant power of recognizing interactions and non-interactions. When the prediction according to the score $p_{i,j;z}$ is incorrect, the second term in Equation (7) might save the prediction.

2.4. Regularization term

The regularization term we introduced in (1) is

$$\theta_0 \| \mathbf{e} \|_1 = \sum_{i<j} e_{ij} \mathbf{w}_s, \tag{8}$$

where $\theta_0 = \mathbf{w}_s$, $\mathbf{w}_s \in R$. When relaxing $e \in \{0, 1\}$ to $e \in [0, 1]$, indeed this regularization term follows the definition of L_1 norms. Hence we can take this term to enfore the learning of sparse human interactions.

2.5. Training

Suppose 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 interaction configuration among people in the image of the k-th training instance, 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:

$$\begin{aligned} &\min \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{l=1}^{n} \xi_l^2 \\ &\text{s.t. } f_0(\hat{\mathbf{a}}^l, \hat{\mathbf{e}}^l; \boldsymbol{\theta}) - f_0(\mathbf{a}^l, \mathbf{e}^l; \boldsymbol{\theta}) \geq \Delta(\mathbf{a}^l, \mathbf{e}^l, \hat{\mathbf{a}}^l, \hat{\mathbf{e}}^l) - \xi_l, \\ &\quad \forall l, \hat{\mathbf{a}}^l \neq \mathbf{a}^l, \hat{\mathbf{e}}^l \neq \mathbf{e}^l, \xi_l \geq 0 \, \forall l. \end{aligned} \tag{9}$$

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

$$\begin{aligned} &\Delta(\mathbf{a}, \mathbf{e}, \hat{\mathbf{a}}, \hat{\mathbf{e}}) = \\ &\frac{1}{m} \sum_{k=1}^{m} \delta(a_k \neq \hat{a}_k) + \frac{2}{m(m-1)} \sum_{i<j} \delta(e_{ij} \neq \hat{e}_{ij}), \end{aligned} \tag{10}$$

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. The problem (9) is convex, which can be solved via the cutting plane algorithm proposed in [16].

2.6. The inference algorithm

Here we discuss the method of solving the inference problem (1) with fixed $\boldsymbol{\theta}$ parameters. Since the problem is NP hard, it is impossible to obtain global solutions in general. One can relax it into a bilinear programming problem (with introducing many auxiliary variables) and solve the relaxation via branch and bound [11]. Instead we solve the inference approximately using an alternating search strategy.

Our algorithm optimize over the labelling space and the interaction structure space in turn in a few iterations. 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:

$$\min_{\mathbf{a}} \sum_{i \in V} \sum_{s \in \mathbb{A}} \theta_{i;s} \mathbb{1}_s(a_i) + \sum_{j<j, \hat{e}_{ij}=1} \sum_{(s,t) \in \mathbb{A}^2} \theta_{i,j;s,t} \mathbb{1}_{s,t}(a_i, a_j), \tag{11}$$

which is solved by the tree-reweighted message passing routine in the OpenGM package [17].

Denoting the current solution of $\mathbf{a}$ by $\hat{\mathbf{a}}$. Submitting $\hat{\mathbf{a}}$ to (1), the inference problem shrinks to

$$\min_{\mathbf{e}} \sum_{j<i} \sum_{z \in \{0,1\}} \theta_{i,j;z} \mathbb{1}_z(e_{ij}) + e_{ij} \theta_{i,j;\hat{a}_i,\hat{a}_j} + \theta_0 e_{ij}. \tag{12}$$

Reorganizing the terms in the objective in (12) we get the following equivalent form:

$$\min_{\mathbf{e}} \sum_{j<i} \left[\theta_{i,j;1} - \theta_{i,j;0} + \theta_{i,j;\hat{a}_i,\hat{a}_j} + \theta_0 \right] e_{ij}. \tag{13}$$

Here each e_{ij} has no interaction with 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 ouput of $\theta_{i,j;1} - \theta_{i,j;0} + \theta_{i,j;\hat{a}_i,\hat{a}_j} + \theta_0$ is negative (zero or positive).

3. EXPERIMENTS AND RESULTS

We evaluate our approach on UT and TVHI for human action recognition in natural scene. TVHI is a collection of 300 video clips from 23 different TV shows [18]. It contains five action classes: handshake (HS), highfive (HF), hug (HG), kiss (KS) and no-interaction (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 with respect to the numbers of examples of different action classes. UT [19] contains 120 short videos six action classes: handshake, hug, kick, punch and push. UT includes three asymmetrical actions: kick, punch and push. As done in [12], we add a complementary class for each asymmetrical class. As a result, we have 9 action classes including no-interaction (NO), handshake (HS), hug (HG), kick (KK), be-kicked (BKK), push (PS), be-pushed (BPS), punch (PC), be-punched (BPC).

We compare our method with baselines and the state-of-the-arts. The baselines include 1) *HoG+HoF*: the linear SVM model trained with the HoG+HoF descriptor used by [12]; 2) *Spatial Net*: the spatial net proposed by [1] for action recognition; 3) *Combined+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*: 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:

- *sparse CRF*: the CRF model proposed recently in [12]. Both the interactions and actions are learned separately from data, which is different from the joint learning approach proposed in this paper.

Table 1. Action recognition accuracy (%) on TVHI.

method	no-action	handshake	highfive	hug	kiss	non-interacting	interacting	overall	mean
HoG+HoF	52.4	26.5	25.7	20.0	36.1	–	–	44.5	32.1
Spatial Net [1]	54.7	38.4	35.4	54.8	58.6	–	–	53.6	48.4
Combined+SVM	59.9	41.7	46.0	56.2	66.7	–	–	58.5	54.1
dense CRF+Combined	70.4	32.0	32.0	55.5	**70.0**	0	**100**	47.3	51.4
sparse CRF[12]	82.1	16.0	14.0	30.1	42.7	73.5	52.5	67.1	43.6
latent CRF[9]+Combined	**86.0**	28.3	22.1	45.1	43.1	78.2	48.7	**71.6**	50.2
sparse CRF[12]+Combined	59.9	**46.9**	43.8	58.3	63.8	84.7	62.8	67.5	60.0
joint learning (ours)	57.9	46.0	**46.9**	**63.1**	67.9	**86.2**	67.4	68.1	**62.2**

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 [1]	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+Combined	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[12]	**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[9]+Combined	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[12]+Combined	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 learning (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**

• *sparse CRF+Combined*: 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.

• *latent CRF+Combined*: the approach proposed in [9], which treats interaction configuration as latent variables and estimates actions and interactions simultaneously. This is different from our approach that learns actions and interactions jointly and explicitly from data. For fair comparison this model also utilizes combined descriptors and head orientations.

3.1. Evaluate 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 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 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. This is because that latent CRF performs better on the *no-action* class, which dominates the TVHI dataset with respect to the number of examples. Note that latent CRF performs much worse on all other classes. Hence our approach admits much unbaised results across different classes compared against latent CRF.

3.2. Evaluate on UT

For UT we use the same evaluation protocol as TVHI. Results are shown in Table 2. Again the proposed approach outperforms the baseline significantly with respect to action classification, and is competitive compared with the state-of-the-arts.

4. CONCLUSION

We have presented an approach to learn the configuration of human interactions and their action labels in a joint framework. Our formulation learns both actions and human interactions explicitly in a supervised manner, using both deep-neural-network features and high-level contextual information. We used a max-margin-style training method to learn model parameters, and proposed an efficient optimization algorithm to solve the relevant inference problem. The proposed joint learning method achieves superior results on both action and interaction recognitions, as evidenced by comparison with baselines and the state-of-the-arts.

5. ACKNOWLEDGEMENT

Zhenhua Wang was partially supported by the 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). Qiu Guan was supported in part by Public welfare project of Zhejiang provincial Department of Science and Technology (2015C33073).

6. REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014. 1, 2, 3, 4

[2] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. van Gool, "Temporal segment networks: towards good practices for deep action recognition," *arXiv preprint arXiv:1608.00859*, 2016. 1

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013. 1

[4] E. Park, X. Han, T. Berg, and A. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *Winter Conference on Applications of Computer Vision (WACV)*, 2016. 1

[5] T. Du, L. Bourdev, R. Fergus, and L. Torresani, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015. 1

[6] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR*, 2011. 1

[7] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition," in *CVPR*, 2009. 1

[8] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 210–220, 2006. 1

[9] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, 2012. 1, 4

[10] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*, MIT Press, 2009. 1

[11] Z. Wang, Q. Shi, C. Shen, and A. van den Hengel, "Bilinear programming for human activity recognition with unknown mrf graphs," in *CVPR*, 2013. 1, 3

[12] Z. Wang, S. Liu, J. Zhang, S. Chen, and Q. Guan, "A spatio-temporal crf for human interaction understanding," *IEEE Trans. Circuits Syst. Video Technol.*, 2016. 1, 3, 4

[13] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2441–2453, 2012. 1

[14] H. Wang and C. Schmid, "Action recognition with improved trajactories," in *ECCV*, 2013. 2

[15] B. Ahn, J. Park, and I. Kweon, "Real-time head orientation from a monocular camera using deep neural network," in *ACCV*, 2015. 2

[16] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, no. 2, pp. 1453–1484, 2006. 3

[17] B. Andres, T. Beier, and J. H. Kappes, "Opengm," http://hciweb2.iwr.uni-heidelberg.de/opengm/index.php, 2015. 3

[18] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High five: recognising human interactions in tv shows," in *BMVC*, 2010. 3

[19] M. S. Ryoo and J. K. Aggarwal, "UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA)," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010. 3