Multi-task Recurrent Neural Network for Immediacy Prediction

Xiao Chu   Wanli Ouyang   Wei Yang   Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
xchu@ee.cuhk.edu.hk   wlouyang@ee.cuhk.edu.hk   xgwang@ee.cuhk.edu.hk

Abstract

In this paper, we propose to predict immediacy for interacting persons from still images. A complete immediacy set includes interactions, relative distance, body leaning direction and standing orientation. These measures are found to be related to the attitude, social relationship, social interaction, action, nationality, and religion of the communicators. [1] *A large-scale dataset with $10,000$ images is constructed, in which all the immediacy cues and the human poses are annotated. We propose a rich set of immediacy representations that help to predict immediacy from imperfect 1-person and 2-person pose estimation results. A multi-task deep recurrent neural network is constructed to take the proposed rich immediacy representations as the input and learn the complex relationship among immediacy predictions through multiple steps of refinement. The effectiveness of the proposed approach is proved through extensive experiments on the large-scale dataset.*

1. Introduction

The concept of immediacy was first introduced by Mehrabian [18] to rate the nonverbal behaviors that have been found to be significant indicators of communicators' attitude toward addressees. In [18], several typical immediacy cues were defined: touching, relative distance, body leaning direction, eye contact and standing orientation (listed in the order of importance). A complete set of immediacy cues defined in this work are shown in Fig. 1. These cues are important attributes found to be related to the inter-person attitude, social relationship, and religion of the communicators [17, 36, 12]. Immediacy cues report the communicators' attitude which is useful in building up social networks. With vast data available from social networking sites, connections among people can be built up automatically by analyzing immediacy cues from visual data. Second, these immediacy cues are useful for existing vision tasks, such as human pose estimation [38, 32], social relationship, social role [27], and action recognition [16]. The immediacy cue "touch-code" is the same as interaction recognition and has been recognized by our society [37, 13, 23] for a long time. However, a complete dataset providing all the immediacy cues is absent. In addition, there is only little research on immediacy analysis from the computer vision point of view.

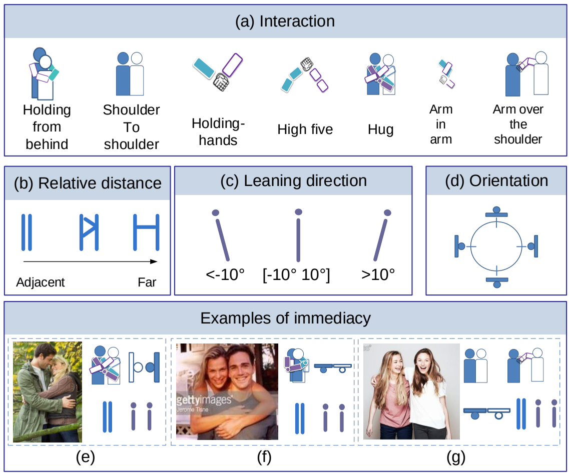

Figure 1. The tasks of immediacy prediction and three examples. Detailed definitions of immediacy cues can be found in Sec. 3

In order to predict immediacy, it is natural to use the information from 1-person pose estimation [38] and 2-person pose estimation, which was called touch-code in [37]. However, touch-code and single person pose estimation are imperfect. Especially, when people have interaction, inter-occlusion, limb ambiguities, and large pose variation inevitably occur. These cause the difficulty in immediacy prediction. On the other hand, interacting persons provide extra representations that motivate our work.

First, there are extra information sources unexplored when persons interact. Since both 1-person or 2-person pose estimation are imperfect, extra information sources, *i.e.*, overlap of body parts, body location relative to two persons' center, and consistency between 1-person and 2-person estimation, are helpful for immediacy prediction as well as addressing pose estimation errors. As an example for overlap of body parts, when all of person A and person

[1] The dataset can be found at http://www.ee.cuhk.edu.hk/~xgwang/projectpage_immediacy.html

B's body parts have large overlap, it is more likely that A is holding B from behind but less likely that A is shoulder-to-shoulder with B. Regarding relative location to two persons' center, if person A's right shoulder and persons B's left shoulder are close to the two persons' center, shoulder-to-shoulder is more likely to happen. As for consistency between 1-person and 2-person estimation, when 1-person and 2-person estimation results conform with each other, they are more reliable than results that do not. Therefore, we include these extra information sources and jointly consider both 1-person and 2-person pose information by integrating them with a deep model, so that these information sources validate each other's correctness in conforming with the global pattern of human interactions as well as other immediacy cues.

Second, immediacy cues are correlated. For instance, shoulder-to-shoulder only happens when two persons stand close to each other (distance) and side-by-side (standing orientation). Existing works treat interactions independently but do not explore the correlation mentioned above. We hence construct a multi-task recurrent neural network (RNN) to learn the correlations among immediacy as well as the deep representations shared by all the immediacy cues. RNN fits this problem because it can iteratively refine the coarse outputs from a deep model through multiple steps in the forward pass and pass estimation errors back to the deep model during training. Therefore, the whole model can be trained end-to-end with back propagation.

Our work contributes in the following three ways.

- Propose the immediacy prediction problem, and build a large-scale dataset that contains 10,000 images. It has rich annotations on all immediacy measures and human poses.
- Rich immediacy representations by taking extra information sources into consideration, *i.e*., overlap of body parts, body location relative to two persons center, and consistency between 1-person and 2-person estimation. For predicting immediacy, a unified deep model is used for capturing the global interaction patterns from the proposed immediacy representations.
- Construct a multi-task deep RNN to model complex correlations among immediacy cues as well as 1-person and 2-person pose estimation. The recurrent neural network is used for refining coarse predictions through multiple steps. We prove that by jointly learning all tasks of predicting a complete set of immediacy cues with deep RNN, the performance of each task can be boosted dramatically, and pose estimation can be improved as well.

2. Related Work

Pose estimation is an input of our approach. Remarkable research progress in pose estimation has inspired our work and is the base of our work. Both holistic models [28, 11, 19, 20, 33] and local models [31, 8, 39, 6, 9, 30, 24, 10, 26, 14, 29, 35, 25, 1, 21, 4] were used for estimating the pose of a single person. For multi-person pose estimation, some approaches used occlusion status [7] and spatial locations [37, 15, 3] of body parts as pairwise constraints. Existing works either considered only single person pose estimation [31, 34, 8, 39, 6, 9, 30, 24, 2, 26, 14, 29, 35, 25, 1, 21] or multi-person pose estimation [37, 13], while our model jointly takes 1-person and 2-person pose estimation as the input. Our work targets on predicting immediacy, although we do find prediction on immediacy can improve pose estimation.

Interaction pattern is called proxemics and was estimated using deformable part based models [13] and flexible mixtures of part models in [37]. It was estimated from videos using motion in [23]. Human distance was used in [3] for estimating social activities such as crowd and speaker-audience. 2-person pose estimation was used for interaction estimation in [37]. We differ in three aspects. First, existing works only consider single factors like interaction or distance, while our work is the first towards complete study on estimating all immediacy cues, including interaction, relative distance, body leaning direction and standing orientation, and learning them jointly. Second, extra information sources, *i.e*., overlap, relative location to center, and consistency between 1-person and 2-person pose estimation, were not explored in these works. Third, existing works modeled the interaction types separately, while we construct a RNN to jointly learn their relationships.

Multi-task learning [40] have been used to directly model the relationships among correlated tasks. Immediacy prediction is a highly non-linear mapping function of the information from pose estimation, and therefore non-linear deep representation is desirable. The correlations among immediacy cues are also complex. In our work, such deep representation and immediacy relationship representation are jointly learned with an end-to-end deep multi-task RNN model. Razvan Pascanu *et al*. [22] provided many structures of deep RNN, which are usually used to model sequential data. Zheng *et al*. [41] showed that the iterative belief propagation in CRF can be approximated with RNN. These motivate us to use RNN to model the complex correlations among tasks. The predictions on multiple tasks at the previous step are used as the input for the network at the next step. They are coupled with the original input data through multi-layer nonlinear mapping to refine the prediction. As the refinement goes through more steps in RNN, higher nonlinearity can be modeled.

3. Immediacy Dataset

We construct a large-scale dataset for the proposed immediacy prediction problem. It consists of 10,000 images. Data are collected from four major sources: Getty Images website, photos of celebrities, movies and drama series.

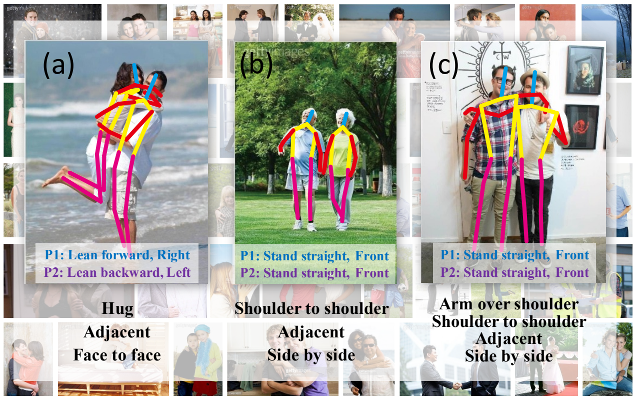

Figure 2. Examples of images in our dataset and their annotations. P1 represents the person on the left and P2 represents the person on the right.

Previous related benchmark datasets: the WeAreFamily dataset [7] is used for multiple people pose estimation; The TVHI dataset (video) [23] and TVHI+ dataset (video) [13] are used for interaction classification. The proxemics dataset [37] provides labels for both human pose and interaction classes, but its scale is relatively small and the data are lack of diversity (only including persons in front view). Quantitative comparison is listed in Table 1. The comparison with TVHI is based on the frames used in [23].

Dataset	Annotation	Training	Test
Proxemics [37]	pose and interaction	300	289
WeAreFamily [7]	pose	350	171
TVHI [23]	interaction	599	714
TVHI+ [13]	interaction	1654	1566
Immediacy (ours)	interaction, pose and posture attributes	7500	2500

Table 1. Quantitative comparison of related datasets.

We provide multiple annotations for every image. Definitions of immediacy cues are shown in Fig. 1 and examples of annotated images are shown in Figure 2. Details about the annotations are listed as follows:

1. Pose: we annotate the pose of each person. For images containing lower body parts, we annotate the full pose. For the rest images, we only annotate the upper bodies. The visibility of each body joint is also annotated.
2. Interaction: we define 7 kinds of frequently appearing interactions as listed in Table 2. We annotate the interaction for each pair of persons. A pair of persons can be annotated with more than one kind of interactions. For instance, in Figure 2 (b), the man is "holding hands" and "shoulder to shoulder" with the woman.
3. Relative distance: the physical distance separating one person from another.It is quantified into three levels according to the scale of half arm. If two persons are standing closely, *i.e.*, within the reach of half arm, we label this distance as "adjacent". If they are farther than the reach of one arm, the distance is labeled as "far". The rest is "near".
4. Leaning direction: we annotate the leaning direction of each person as leaning forward to the other person, leaning backward or stand straight.
5. Orientation: whether a person is facing the camera (front view), facing left, facing right or facing back.
6. Relative orientation: relative body orientation of paired persons, *i.e.*, face to face, 90 degree or side by side.

1	Holding from behind	HB	869	Proxemics [37]
2	Hug	HG	2221	TVHI [23]
3	Holding hands	HH	1718	TVHI, Proxemics
4	High five	HF	613	TVHI
5	Arm over the shoulder	AS	3182	Proxemics
6	Shoulder to shoulder	SS	3453	Proxemics
7	Arm in arm	AA	1046	Proxemics

Table 2. Seven classes of interactions. The second column is the name for each class and the third column is its abbreviation we shall use in the following sections. The number of images under each kind of interactions is listed in the forth column. The fifth column shows which dataset previously defined such kind of interaction.

Our dataset is more challenging and more suitable for practical use in the following aspects: the age of characters varies from infancy to adulthood; the postures of human are diverse, including lying, sitting and standing; all kinds of viewpoints are included (front view, side view and back view); frames extracted from videos are not consecutive.

4. Overview

As shown in Figure 3, the overview of our approach for immediacy prediction is as follows:

1. 1-person pose estimation [38] is used for extracting the unary representation.

2. 2-person pose estimation approach [37] is used for obtaining pairwise feature representations. The unary representation and pairwise representation are basic representations that are extracted from existing approaches.

3. Based on the 1-person and 2-person pose estimation results, we proposed new representations. These new representations capture distinctness between a pair of poses, relative locations of a pair of poses and consistency between 1-persons pose estimation and 2-person pose estimation.

4. All representations mention before together called immediacy representations, which is used as the input of the multi-task deep RNN.

5. The multi-task deep RNN predicts the immediacy cues and poses. Since there are many candidates of Ψ, the prediction with the largest confidence in a local region is selected as the final results.

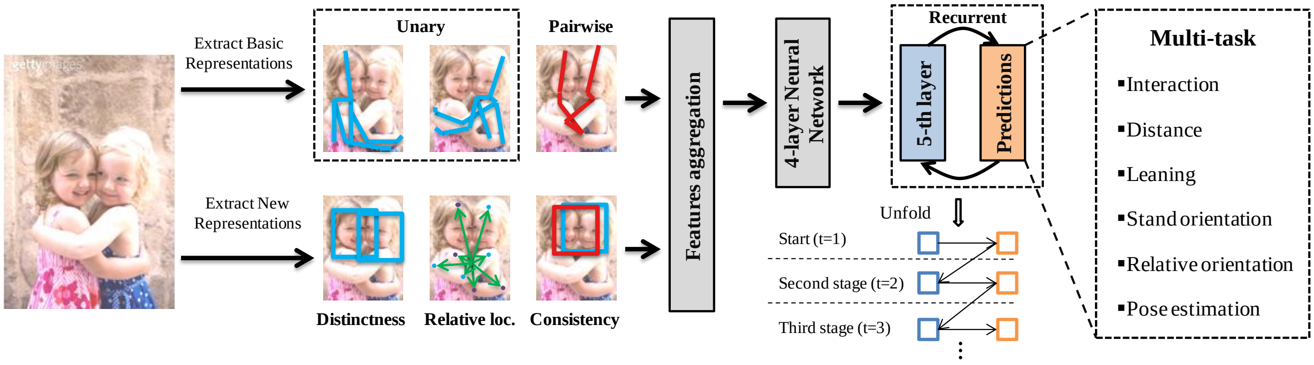

Figure 3. Overview of our approach for immediacy prediction and pose estimation.

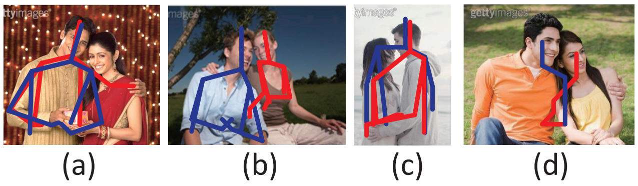

Figure 4. Failure cases of 1-person pose estimation and 2-person pose estimation. For pose estimation, the double counting (a), missing body joints (b) and miss alignment of body joints (c) are major problems, especially when multiple persons interact. For 2-persons pose estimation, false positive on wrong class frequently occurred.

5. Rich immediacy representations

5.1. Basic representations

Unary representation Ψ_u describes the features of a single person's pose. A classical pose estimation approach [39] is used for feature extraction. In this approach, the pose of a person is modeled by a tree-structured graph, with each node in the graph representing a body joint. Ψ_u can be expanded into the following components: $\Psi_u = [\Psi_u^a, \Psi_u^d, \Psi_u^m]$. Ψ_u^a is the appearance score for each body joint, which indicates the correctness of part appearance. Ψ_u^d is the relative location between one part, e.g. hand, and another part, e.g. elbow. Ψ_u^d is normalized by the human head's scale. This term describes the articulation of a pose. Ψ_u^m is the mixture type of each body joint.

Pairwise representation Ψ_p mainly captures information when two persons interact. Previous work on recognizing proxemics [37] restricted the way of describing interactions to touch-codes. Inspired by their work, we train multiple models to capture the touch-codes in 7 kinds of interactions. Ψ_p is composed of Ψ_p^a, Ψ_p^m and Ψ_p^d. The meaning of each term is same with the corresponding terms inΨ_u. Ψ_p^a is the appearance score of each body joint, Ψ_p^m is the mixture type and Ψ_p^d is the relative location.

The models employed to extract pose features from images are imperfect. The examples in Figure 4 show the major problems existing in current approach. These problems lead to the unreliability of the basic feature representations Ψ_u and Ψ_p. Therefore, extra representations shall be introduced in the following section to assist immediacy estimation.

5.2. New representations

Distinct pose representation Ψ_{ov} measures the similarity between a pair of poses from single person pose estimation. In pose estimation, a bounding box is defined for each body joint to extract its visual cue. The bounding box for the pth body joint is called the pth part-box and denoted by box_p. The overlap of each part ov_p is defined by the intersection over the union as follows:

$$ov_p = \frac{\cap(box_p^1, box_p^2)}{\cup(box_p^1, box_p^2)}, \tag{1}$$

where box_p^1 and box_p^2 are the p-th part-box for the first person and the second person respectively. Only the part-boxes for the same body joint of paired persons are considered in this representation. In our framework, large amount of overlaps for many body parts can be accepted by interaction classes such as "holding from behind" and "hug", and rejected by interaction classes such as "holding hands". $\Psi_{ov} = [ov_1, ..., ov_P]$ also implicitly improves non-maximum suppression (NMS). One body part could generate two part-boxes during pose estimation, and they could be wrongly interpreted as coming from two persons when modeling interaction, instead being merged into one by NMS. The representation Ψ_{ov} can identify such potential cases and help to address pose estimation errors in the higher layers of the neural network.

Relative location representation, denoted by Ψ_l, captues the relative locations of poses from 1-person pose estimtation to their center.

$$l_p^k = ([x_p^k, y_p^k] - \frac{1}{KP}\sum_k^K \sum_p^P [x_p^k, y_p^k])/p_{scale}, \tag{2}$$

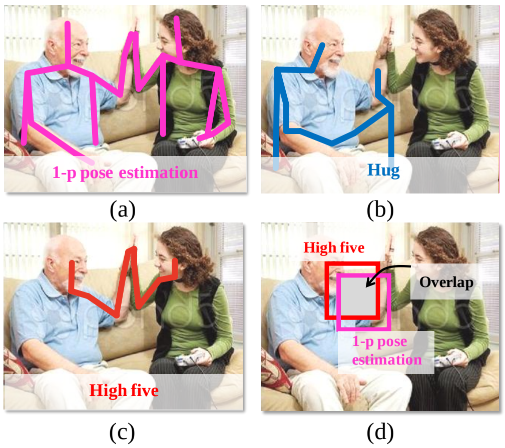

Figure 5. Measuring the consistency between 1-person pose estimation (a) and 2-person pose estimation (b)(c). The consistency is measured by calculating the overlap between partbxes (d).

where $[x_p^k, y_p^k]$ is the location of the p-th part for the k-th person. Since the scale of images and scale of persons vary a lot across datasets, we choose the center of two paired persons as the origin of coordinate plane, and normalize the location of each body part by the scale p_{scale} of bounding boxes for the body parts. In our approach, $K = 2$ and $P = 23$. Relative location representation is useful for prediction of distance and interaction. For example, when the relative locations to the center are large for most parts, the distance should be far and the interaction class is more likely to have "high-five" but less likely to have "holding from behind".

Consistency representation, denoted by Ψ_m, measures whether 1-person pose estimation matches with 2-person pose estimation results. To be more specific, the head location predicted by 1-person pose estimation model should be close to the head location predicted by 2-person pose estimation model. $\Psi_m = [\Psi_m^1, \ldots, \Psi_m^n, \ldots, \Psi_m^N]$, where $\Psi_m^n = [ov_{1,m}^n, \ldots, ov_{j,m}^n, \ldots ov_{j,m}^n]$ and $ov_{j,m}^n$ is the overlap of the jth part-box between the 1-person pose estimation result and the 2-person pose estimation result of type n. N denotes the number of body parts estimated in the 2-person pose estimation. In Figure 5, two persons have high five. As the input of our model, 2-person pose estimation provides multiple candidates, which lead to different prediction scores on interaction categories. For example, the candidate in Figure 5 (b) generates a high score on "hug", while the candidate in Figure 5 (c) predicts "high five". By checking this consistency representation, we find that the interaction of "high five" has more overlap with the 1-person pose estimation results as shown in Figure 5 (a) and (d). Therefore, this consistency representation could help the 1-person pose estimation and 2-persons pose estimation to validate each other.

In summary, we use features from 1-person and 2-person pose estimation. Instead of simply concatenating them, we propose three new representations for predicting immediacy: Ψ_u^{ov} measures the similarity of the pair of poses from 1-person pose estimation, Ψ_u^l describes the relative body location to the person center for the pair of poses from 1-person pose estimation, and Ψ_m measures the consistency between 1-person and 2-person pose estimation.

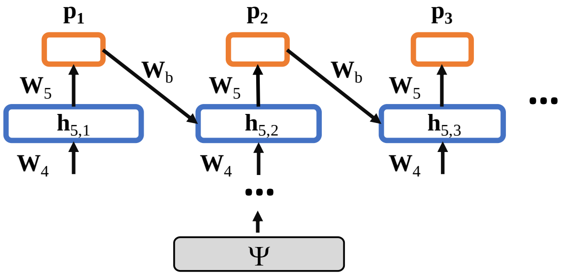

Figure 6. Multi-task RNN structure.

6. Modeling multi-task relationships with RNN

The immediacy cues are correlated with each other strongly. Their complex relationships cannot be well captured with a single network. Our idea is to replicate the network and refine the predictions through multiple steps as shown in Figure 6. The coarse prediction from the first network is used as the input of the hidden layer of the second network, which also takes the original data as input in the bottom layer. This process can be repeated through multiple steps. As the number of the steps increases, more complex relationship could be modeled.

6.1. Multi-task deep RNN

Denote the concatenation of the representations introduced in Section 5 as Ψ. A 4-layer neural network is built to learn deep representations from Ψ as follows:

$$\mathbf{h}_{1,t} = f(\mathbf{W}_1 \Psi + \mathbf{b}_1), \tag{3}$$

$$\mathbf{h}_{l,t} = f(\mathbf{W}_l \mathbf{h}_{l-1,t} + \mathbf{b}_l), \text{ for } l = 2, \ldots 4, \tag{4}$$

where $\mathbf{h}_{l,t}$ denotes the l-th layer. $f(\cdot)$ is the element-wise non-linear activation function. $\mathbf{W}_l$ contains the weight parameters and $\mathbf{b}_l$ contains the bias parameters.

Denote the prediction on immediacy cues at step t as $\mathbf{p}_t$. After $\mathbf{h}_4$ being extracted with Equation (4), RNN models the relationship among immediacy cues as follows:

$$\mathbf{h}_{5,t} = f(\mathbf{W}_5^T \mathbf{h}_{4,t} + \mathbf{W}_b^T \mathbf{p}_{t-1} + \mathbf{b}_5) \tag{5}$$

$$\mathbf{p}_t = f(\mathbf{W}_{cls}^T \mathbf{h}_{5,t} + \mathbf{b}_{cls}), \tag{6}$$

where $\mathbf{W}_5$ is the weight from $\mathbf{h}_{4,t}$ to $\mathbf{h}_{5,t}$, $\mathbf{W}_b$ is the weight from $\mathbf{p}_{t-1}$ to $\mathbf{h}_{5,t}$, $\mathbf{W}_{cls}$ is used as the prediction classifier, $\mathbf{b}_5$ and $\mathbf{b}_{cls}$ are bias terms. At step t in (5), hidden variables in $\mathbf{h}_{5,t}$ are updated at each step using its hidden variables in $\mathbf{h}_4$ and the predicted immediacy $\mathbf{p}_{t-1}$ at the previous time step $t-1$. The predicted immediacy $\mathbf{p}_t$ in (6) is obtained from the updated hidden variables in $\mathbf{h}_{5,t}$.

There are other choices of RNN structures, such as 1) directly connecting $\mathbf{p}_{t-1}$ with $\mathbf{p}_t$ instead of $\mathbf{h}_{5,t}$; or 2) connecting $\mathbf{h}_{5,t-1}$ (instead of $\mathbf{p}_{t-1}$) to $\mathbf{h}_{5,t}$. Experiments show that the structure in Figure 6 is most suitable for our problem and dataset. In option 1), there is only one-layer nonlinear mapping between $\mathbf{p}_{t-1}$ and $\mathbf{p}_t$ and hence, it cannot well model complex relationships. In option 2), the influence from the previous predictions to the current predictions is transmitted by hidden variables $\mathbf{h}_{5,t-1}$, which is more indirect and hard to learn given a limited dataset.

6.2. Learning

The i-th sample for the c-th immediacy cue is denoted as $(\Psi_{(i)}, y^c_{(i)})$, where $y^c_{(i)}$ is the label for c-th immediacy cue. The parameter set $\Theta = \{\mathbf{W}_*, \mathbf{b}_*\}$ in (3)-(6) is learned by back propagation using the following loss function:

$$\arg\min_{\Theta} \sum_i \sum_c \lambda^c y^c_{(i)} \log p(y^c_{(i)} | \mathbf{\Psi}_{(i)}; \Theta) + \sum_c \|\mathbf{w}\|_2^2, \quad (7)$$

where $\mathbf{w}$ is the concatenation of all elements in $\mathbf{W}_*$ into a vector.

6.3. Analysis

The hidden variables at higher layers (with larger l) progressively extract more abstract feature representations. The hidden variables in $\mathbf{h}_{5,t}$ summarize the correlations of immediacy cues. The immediacy cues can be mutually consistent or exclusive.

When two immediacy cues are mutually consistent, the existence of one cue reinforces the confidence on the existence of another cue. For example, “shoulder to shoulder” often happens together with “arm in arm”. Once “arm in arm” appears, the “shoulder to shoulder” has its prediction confidence raised.

If two immediacy cues are mutually exclusive but confident prediction scores are assigned to both of them in the preliminary prediction stage, then there is a conflict between the predictions. The hidden variables in $\mathbf{h}_{5,t}$ have access to the information of the lower layer $\mathbf{h}_{4,t}$ as well as the prediction results $\mathbf{p}_{t-1}$ in the previous step. $\mathbf{h}_{5,t}$ notices this conflict by using information from both $\mathbf{h}_{4,t}$ and $\mathbf{p}_{t-1}$ in order to make a decision on which conflicted prediction is wrong. For example, “holding hands” is mutually exclusive to “high five”. In Figure 7, the preliminary prediction $\mathbf{p}_{t-1}$ has unreasonably high responses to both “holding hands” and “high five”. $\mathbf{h}_{5,t}$ finds this conflict from $\mathbf{p}_{t-1}$. Then it is able to figure out that “high five” is correct but “holding hands” is wrong through nonlinear reasoning from $\mathbf{h}_{4,t}$ and $\mathbf{p}_{t-1}$. The response of “holding hands” is finally suppressed.

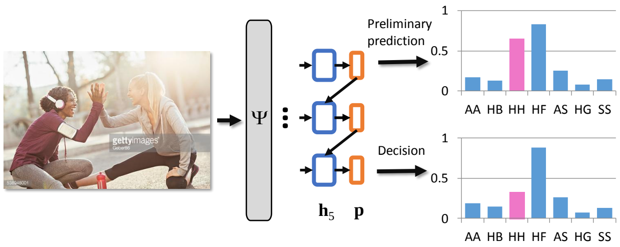

Figure 7. Illustration of our proposed multi-task RNN. The image on the left is the input. Predictions are on the right. The horizontal axis shows six classes of interactions, and the vertical axis is the true positive rate of specified interaction. The preliminary predictions on 7 classes of interaction are reported on the top, while the refined predictions are reported at the bottom.

7. Experiment

We mainly use the immediacy dataset introduced in Section 3 for training and testing. In training stage, negative images from INRIA [5] are used. In the training and testing stage, it is assumed that the bounding boxes of the two interacting persons are given, so that the algorithm knows which persons are targets of interest among people in an image. On our dataset, we compare the results on 7 classes of interactions and the other immediacy cues, *i.e.*, relative distance, body leaning direction and standing orientation. We also briefly evaluate pose estimation results. Immediacy factors are measured between two specified persons. The overlap between the provided person bounding boxes and prediction candidates is calculated. If the overlap is smaller than 0.3, the candidate will be ignored at the evaluation stage.

7.1. Predictions on interactions

Method	**HB**	**AA**	**HH**	**HF**	**AS**	**SS**	**HG**	**mAP**
svm+Ψ_u	24.8	18.6	32.5	37.6	51.0	70.4	50.5	40.7
proxemics [37]	45.5	59.6	52.4	83.7	59.9	80.1	61.1	63.2
deep+Ψ_p	43.5	60.0	64.0	82.7	67.7	74.9	68.4	65.9
deep+Ψ_u+Ψ_p	46.6	62.3	66.8	82.2	69.6	86.3	66.6	68.6
deep+Ψ	47.5	68.6	73.0	83.7	77.6	89.8	78.1	74.0
d+Ψ+joint	56.8	70.2	73.8	86.8	79.0	89.0	79.8	76.7
d+Ψ+j+RNN	**64.3**	**73.6**	**74.5**	**88.4**	**82.3**	**90.2**	**82.3**	**79.8**

Table 3. Interaction prediction results. The second row shows the results on the existing method in [37].

In the implementation of our approach for interaction prediction, we use the bounding box covering both persons to reject interaction candidates. For prediction of other immediacy measures and pose, the bounding boxes of two persons are used to reject pose candidates. The evaluation criteria we use is mean average precision (mAP), the same as [37]. Experiment results are shown in Table 3.

Investigation on using single representation. The *svm*+Ψ_u in Table 3 denotes the results of only using unary representation Ψ_u from 1-person pose estimation. SVM is used as the classifier. Because of inter-occlusion and limb ambiguities, it only has mAP of 40.7%. The *proxemics* in Table 3 denotes results of using the approach in [37]. We retrain their models on our dataset with their released code. This approach also provides us the pairwise representation Ψ_p. Proxemics estimation is better than directly using the unary representation for paired persons.

Investigation on using multiple representations. In this set of experiments, we use multi-layer deep models for evaluation. Each task is treated separately, aiming at evaluating the effectiveness of our proposed immediacy representations $\Psi = [\Psi_u, \Psi_p, \Psi_{ov}, \Psi_l, \Psi_m]$. deep+$\Psi_p$ denotes the result that only uses Ψ_p (pairwise representation in Sec. 5) as the input to the deep model, with mAP 65.9%. *deep*+Ψ_u+Ψ_p denotes the result that only uses the basic representations Ψ_u (unary representation) and Ψ_p (pairwise representation) as the input to the deep model. The mAP for this combination is 68.6%. Simply adding the single poses to Ψ_p could not gain obvious improvement. *deep*+Ψ denotes usesour proposed representations Ψ_{ov}, Ψ_l and Ψ_m as input, modeling the input features with deep model. Compared to the result in *deep*+Ψ_u+Ψ_p, large improvement (6.6%) is gained by employing the new representations Ψ_{ov}, Ψ_l, Ψ_m. Therefore, the effectiveness of the newly proposed representations are proved. All the results above are learned independently by using only single interaction class for supervision.

Investigation on learning correlations among interactions. The methods evaluated in this set of experiments are all based on the full set of immediacy representations Ψ and deep models. The *d*+Ψ+*joint* in Table 3 shows the results when all immediacy cues are jointly predicted without RNN for refining predictions. With joint learning of all the cues, the mAP of *d*+Ψ+*joint* gets 2.7% improvement than single task learning, *i.e.*, deep+Ψ. Among all the tasks, the performance of interaction "holding from behind" is dramatically improved from 47.5% to 56.8%. With extra supervision provided by other immediacy tasks, better deep representations can be learned for "holding from behand". However, the overall gain of *d*+Ψ+*joint* compared with *d*+Ψ is not obvious. Our proposed RNN in Figure 6, denoted by d+Ψ+j+rnn, propagates the prediction $\mathbf{p}_{t-1}$ at step $t-1$ to $\mathbf{h}_{5,t}$ at the next step. Experimental results show that it outperforms all the previous methods by 3% in overall mAP. Directly utilizing the predictions at previous step, which are more summarized information and more directly influenced by supervision, could better assist the predictions in the next time step.

7.2. Predictions on relative distance, body leaning direction and orientation

Table 4 shows the results on predicting relative distance, body leaning direction and orientation. The denotations are the same as those in Table 3. The first row *svm*+Ψ_u is the result using only Ψ_u (the representations from 1-person pose estimation) as the input to train SVM, and the mean accuracy is 66.1%. The second row *svm*+Ψ shows the results using the full set of proposed immediacy representations Ψ as the input of SVM. No improvement is achieved because the limitation of SVM in processing such nonlinear relationship. Replacing SVM with the deep model, using the full set of the immediacy representations Ψ, and learning each task independently, the predictions have a mean accuracy of 68.8%, denoted by *deep*+Ψ in Table 3. When jointly learning all the tasks, we find that 3/4 tasks get better results but the task of learning human body absolute orientation fails. This is consistent with our assumption that simply using deep model without stage by stage refinement is not reliable. Our proposed multi-task RNN beats all the previous results on all the tasks, with the highest mean accuracy being 74.1%.

method	dist	ori	lean	rel ori	mean
svm+Ψ_u	78.6	62.1	74.3	49.6	66.1
svm+Ψ	76.1	61.9	74.2	48.8	66.2
deep+Ψ	80.6	68.9	74.8	50.9	68.8
d+Ψ+j	81.2	63.8	75.7	55.0	68.9
d+Ψ+j+rnn	**82.7**	**72.3**	**76.5**	**64.9**	**74.1**

Table 4. Accuracies on other immediacy prediction tasks. All these tasks are multi-class classification problem. We use accuracy of each task as evaluation criteria.

7.3. Pose estimation

Pose estimation is the basic task in our framework. It assists learning immediacy cues. In the process of pursuing better performance of immediacy prediction, we find that the pose estimation result is improved as well. For pose estimation, the percentage of correct keypoints (PCK) proposed in [38] is used as the evaluation criteria. An estimation of body part is defined as correct if it falls within $\alpha \max(h, w)$ pixels of the ground-truth body part location. Here, h and w are the height and width of the bounding box of the upper body respectively. α is the threshold controlling how close the part location to its ground truth location should be. In our experiment, $\alpha = 0.2$. We also use ground truth upper body bounding box to select candidates. All the settings are exactly the same as [39].

Experimental results are reported in Table 5. The *single* in the table is the result of directly applying 1-person pose estimation. The previous approach in [21] used deep model for single person pose estimation, denoted by *deep+1P[21]* in Table 5. Their approach has mean PCK 48.70 on our dataset. The performance is slightly improved compared with single person pose estimation[39]. The result in *deep*+Ψ_u denotes the result of paring two single pose estimates as input and train a deep model to learn the relationship between a pair of poses. The performance is largely boosted by this implementation. The mean PCK of this approach is 51.66. Under the frame work of *deep*+Ψ_u, this improvement is mainly from the idea of checking consistence of two related poses. If one persons is reaching out his hand, another person tends to hold his hands. Based on the $deep + \Psi_u$, the addition of Ψ_p, denoted by *deep*+Ψ_u+Ψ_p, has mean PCK 51.72. Simple addition of interaction feature does not provide much improvement. But with the newly proposed feature Ψ_{ov}, which calculate the consistence between Ψ_u and Ψ_p, denoted by $deep + \Psi$, the mean PCK is further improved to 53.33.

The result in *svm*+Ψ utilizes all the proposed immediacy representations Ψ to train a svm. Mean PCKs get a slightly improvement by 1.5%. In the joint learning process, the pose estimation result is reported in *d*+Ψ+*joint*. An improvement in the mean PCK by 1% is observed, comparing to only learn the pose estimation task, *i.e.*, deep+Ψ. Finally, joint learning with our proposed RNN further improves the mean PCK by 1.9% compared with deep model without RNN, *i.e.*, deep+Ψ.

method	head	shd.	elbow	wrt.	hand	tor.	mean
single[39]	69.5	63.0	42.6	31.8	29.0	43.9	46.96
deep+1P[21]	67.7	61.3	46.4	35.4	32.5	48.9	48.70
deep+Ψ_u	68.2	71.4	48.5	37.2	34.0	50.3	51.66
deep+Ψ_u+Ψ_p	68.5	71.5	48.4	37.2	34.4	50.3	51.72
svm+Ψ	70.8	62.0	44.9	33.8	32.0	48.5	48.36
deep+Ψ	78.0	71.4	48.2	37.1	34.5	50.8	53.33
d+Ψ+joint	82.4	72.6	48.9	35.5	32.0	**55.6**	54.50
d+Ψ+j+rnn	**82.5**	**74.6**	**50.1**	**38.8**	**37.1**	55.4	**56.42**

Table 5. Pose Estimation Results. PCK is used as evaluation criteria. “shd.” for shoulder, “wrt.” for wrist and “tor.” for torso.

7.4. Results on the proxemics dataset

We also compare our method on the publicly available proxemics dataset [37] with their method. Since only five classes of interactions defined in the proxemics dataset are the same as our dataset, we only compare the performance on these five classes. The class named ES in proxemics dataset is the same as the class “holding from back”. We retrain our model on the proxemics dataset.

In [37], there are two settings for evaluation: one is given ground truth face bounding boxes and the other is given the detection results of face bounding boxes. Considering that the face ground truth is a too strong prior, we choose to compare with their results based on face detection. The experimental results are shown in Table 6.

	HH	AS	SS	AA	ES	mean
proxe[37]	36.7	28.0	50.0	35.9	33.5	37.5
ours	**41.2**	**35.4**	**62.2**	**43.9**	**55.0**	**46.68**

Table 6. Experiment results on the proxemics dataset.

7.5. Social relationship

Here we target on an simple example to show predictions on immediacy can reveal social relationships. We select five people, *i.e.*, Michelle Obama, Malia Obama, Hillary Clinton, Vladimir Putin, and Barack Obama to form four pairs of celebrities. They have different social relationships, which could be revealed from immediacy. We obtain images from the top ranked images of the Bing image search engine [2] with the queries such as “Barack Obama and Michelle Obama Photos”. Unrelated images are removed. It can be seen from Figure 8 that Barack Obama has many intimate interactions with his wife Michelle Obama and his daughter Malia Obama like arm over shoulder. And we also find that Obama hugs his daughter less than his wife. The histogram for Obama and his family is very different from the histogram for Obama and other politicians. Obama mainly show shoulder-to-shoulder interaction with Hillary and Putin. The other immediacy cues also help. We find that the average standing distance between Hillary and Obama is closer than Putin and Obama. The standing distance between Obma and his family is closer than the distance between Obama and other politicians.

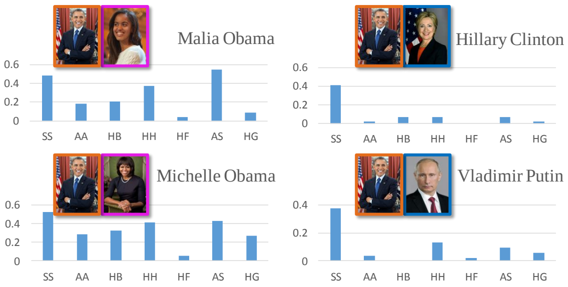

Figure 8. Application. Pink outline indicates family relationship, while blue outline indicates working relationship.

8. Conclusion and Discussion

In this paper, we target on immediacy prediction and construct a dataset, in which annotations of the full set of immediacy cues as well as human poses are provided to facilitate further applications. We propose rich immediacy representations from extra information sources to help immediacy prediction from imperfect 1-person and 2-person pose estimation. A multi-task RNN is proposed to refine coarse predictions through multiple steps and model the complex correlations among immediacy cues. For pose estimation, we observe failure cases when many body parts are occluded or outside the image boundary. Besides, when unusual interaction happens, human pose estimation benefits little from interaction detector.

Acknowledgment: This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK14206114, CUHK14205615, CUHK417011, and CUHK14207814).

[2] http://www.bing.com/images

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: people detection and articulated pose estimation. In *CVPR*, 2009. 2

[3] I. Chakraborty, H. Cheng, and O. Javed. 3D visual proxemics: Recognizing human interactions in 3D from a single image. In *CVPR*, 2013. 2

[4] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014. 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6

[6] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 2

[7] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*. 2010. 2, 3

[8] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *ACCV*. 2013. 2

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, 2005. 2

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 2

[11] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *CVPR*, 2013. 2

[12] E. T. Hall. A system for the notation of proxemic behavior. *American Anthropologist*, 65(5):1003–1026, oct 1963. 1

[13] M. Hoai and A. Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *CVPR*, 2014. 1, 2, 3

[14] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2

[15] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Trans. PAMI*, 34(8):1549–1562, 2012. 2

[16] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1

[17] A. Mehrabian. Relationship of attitude to seated posture, orientation, and distance. *Journal of personality and social psychology*, 10(1):26, 1968. 1

[18] A. Mehrabian. Some referents and measures of nonverbal behavior. *Behavior Research Methods & Instrumentation*, 1(6):203–207, 1968. 1

[19] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV*. 2002. 2

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

[21] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*, 2014. 2, 7, 8

[22] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*, 2013. 2

[23] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *IEEE Trans. PAMI*, 34(12):2441–2453, 2012. 1, 2, 3

[24] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013. 2

[25] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, 2013. 2

[26] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012. 2

[27] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *CVPR*, 2013. 1

[28] B. Sapp and B. Taskar. Multimodal decomposable models for human pose estimation. In *In CVPR*, 2013. 2

[29] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010. 2

[30] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011. 2

[31] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*, 2012. 2

[32] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 1

[33] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2

[34] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, 2013. 2

[35] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008. 2

[36] O. M. Watson and T. D. Graves. Quantitative research in proxemic behavior1. *American Anthropologist*, 68(4):971–985, 1966. 1

[37] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012. 1, 2, 3, 4, 6, 7, 8

[38] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 1, 3, 7

[39] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. PAMI*, 35(12):2878–2890, 2013. 2, 4, 7, 8

[40] Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*, 2012. 2

[41] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. *arXiv preprint arXiv:1502.03240*, 2015. 2