Interaction Relational Network for Mutual Action Recognition

Mauricio Perez, *Member, IEEE,* Jun Liu, and Alex C. Kot, *Fellow, IEEE*

Abstract—Person-person mutual action recognition (also referred to as interaction recognition) is an important research branch of human activity analysis. Current solutions in the field are mainly dominated by CNNs, GCNs and LSTMs. These approaches often consist of complicated architectures and mechanisms to embed the relationships between the two persons on the architecture itself, to ensure the interaction patterns can be properly learned. In this paper, we propose a more simple yet very powerful architecture, named Interaction Relational Network (IRN), which utilizes minimal prior knowledge about the structure of the human body. We drive the network to identify by itself how to relate the body parts from the individuals interacting. In order to better represent the interaction, we define two different relationships, leading to specialized architectures and models for each. These multiple relationship models will then be fused into a single and special architecture, in order to leverage both streams of information for further enhancing the relational reasoning capability. Furthermore we define important structured pair-wise operations to extract meaningful extra information from each pair of joints – distance and motion. Ultimately, with the coupling of an LSTM, our IRN is capable of paramount sequential relational reasoning. These important extensions we made to our network can also be valuable to other problems that require sophisticated relational reasoning. Our solution is able to achieve state-of-the-art performance on the traditional interaction recognition datasets SBU and UT, and also on the mutual actions from the large-scale NTU RGB+D and NTU RGB+D 120 datasets.

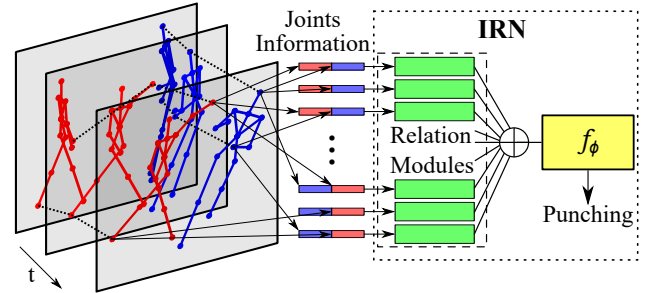*Index Terms*—Interaction Recognition, Pose Information, Relational Reasoning, Skeleton Based.

Fig. 1. Interaction Relational Network (IRN) overview. Joints of both individuals are separately paired up, and then fed to independent relation modules. The pair-wise inferred relationships are then aggregated into a global descriptor used for recognition of the human interaction.

I. INTRODUCTION

RECOGNITION of human interaction (mutual actions) in videos is an important computer vision task, which can help us to develop solutions for a range of applications, such as surveillance, robotics, human-computer interface, content-based retrieval, and so on. Although there have been many works during the past decades [1]–[6], it is still a challenging problem, especially when the videos offer unconventional conditions, such as unusual viewpoints and cluttered background.

Most of the previous works focused on mutual action recognition using RGB videos [1]–[4]. These solutions comprise approaches going from hand-crafted local features to data-driven feature extraction [3], [7]–[10]. Some methods try to implicitly learn the information from the poses of the persons interacting [2], [5]. These solutions often operate over the pixel data, making use of manually annotated regions or poselet labels to focus on these specific regions on the frame or to add extra ad-hoc information. However, they fail to explicitly

model the relationships between the interacting body parts of the involved persons, crucial information for interaction recognition.

With the development of more advanced capturing devices that can also capture depth and extract pose information from the people in the scene (e.g., Microsoft Kinect [11]), a more accurate and valuable description for the human pose can be directly used by new solutions. Besides, very recently, techniques that can estimate poses from regular RGB videos have improved significantly [12]. This allows us to apply reliable pose-based solutions to mutual action recognition in the RGB videos.

The availability of explicit pose representation of the humans in the videos led to a new branch of interaction recognition techniques focusing on this type of data [6], [13]–[17]. These solutions usually involve hand-crafted features and a classification technique that theoretically incorporates the different body-parts' relationships, and then generate a rigid architecture highly dependent on what is believed to be the prior knowledge over the human skeleton structure.

There also exists solutions, using pose information, that target general action recognition [18]–[26], i.e. they include actions performed by a single individual as well. Majority of more recent approaches are based on LSTMs and CNNs, usually ad-hoc network architectures, extra regularization or losses, which tries to better model the spatial and temporal dynamics of the poses during the action. Although these works achieve promising results on general actions, they lack means of leveraging the relationship information between the poses of the two persons when dealing with interactions, underperforming in those cases.

Recurrent models (RNN/LSTM) and convolutional models (CNN/GCN) with carefully designed and complicated architectures have dominated the field of pose-based (skeleton-

M. Perez, J. Liu, and A. C. Kot are with the Rapid-Rich Object Search Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798. E-mail: mauricio001, jliu029, eackot@ntu.edu.sg.

based) action recognition for several years. Instead of using recurrent/convolutional architectures for feature learning, in our paper we propose to view the mutual action recognition task as a body-parts relation inferring and learning problem, and thus design a new architecture to automatically handle this task. Our contribution is based on the powerful Relational Network (RN) [27] for implicitly reasoning over the joints relationships present at human interactions. Although simple, RNs proved to be highly efficient at learning what are the important relationships between pair of objects. Santoro et al. [27] proposed RNs as a simple architecture to deal with problems that require relational reasoning, obtaining state-of-the-art performance on tasks such as visual and text-based question answering, as well as understanding the dynamic of physical systems that included motion data. To the best of our knowledge, no one before have developed a solution based on Relational Network with the purpose of interaction recognition, nor with explicit body information.

On this work we propose Interaction Relational Network (IRN), summarized in Fig. 1. Since our method is the first attempt inspired by the Relational Networks (RNs) for skeleton-based interaction recognition, our first contribution lies on how to logically rearrange our problem components and characteristics in terms that match the high-level concept behind the original RNs. In other words, on how to use the pose information as input, identifying the existing objects (joints) and their properties (coordinates through time), and how to adequately pair these objects, representing the potential relations relevant for interaction recognition. In our new architecture, instead of using pre-defined joints dependency structures as previous methods, we enable our network to infer by itself the existing relations among specific relationships types. We identified two types of relationships, namely, *intra* relations of body joints from the same person and *inter* relations of the joints from different persons, thus the complex relations of different body parts in the person-person interaction sequence are well represented. The relation modules from the same relationship mapping share the same weights, and will infer from the consciously paired up joints which relations exists between them, and which are important for interaction recognition. The relationship mapping, which defines how the joints are paired up, will determine the type of relationships being inferred by that Relation Module (*intra* or *inter*). The description generated from all the relation modules will then be pooled and fed to a module, which will learn how to interpret this arrangement of relations, and classify the interaction accordingly.

Given the specificity of our task and data, we propose two distinct types of relations, however the original RNs contains a single model regardless of relationship type. Thus another contribution in our work is the design of multi-relationship architectures, merging different relation models into a single end-to-end network and leveraging both types of information, therefore leading to a more accurate recognition. We demonstrate the importance of appropriately choosing a fusion architecture, and initializing it with the right models prior to training. Furthermore, we enrich our solution by equipping our new network with structured operations over the object-pair, based on domain knowledge, to extract valuable joint

distance and motion information before reasoning about their relationships. In contrast, the original RNs only contain pair independent operations previously to the relational module. At last, we incorporate an LSTM to our architecture, in order to enable the IRN reasoning over the interactions relationships over a longer duration sequence as well.

We validate our approach through extensive experiments on four different datasets. Two of these are traditional datasets for human interaction recognition: SBU [6] and UT [1]. On which we obtain the state-of-the-art performance, close to perfect accuracy. Seeking to further validate our IRN with more challenging data, we propose to use the large-scale datasets NTU RGB + D [28] and NTU RGB + D 120 [29]. Although these datasets were initially created for benchmarking general human action recognition, they contain many classes of mutual actions, in fact more than SBU and UT. Additionally, they have much more samples per class and adverse evaluation protocols. Even though the NTU RGB + D datasets are more difficult, the IRN still outperforms the previous works.

Our contributions can be summarized in the following manner:

- A novel approach to Human Interaction Recognition, using the different body parts from the pose information as independent objects, and modeling their relationship pairwise.
- We design effective ways to fuse different types of relationships, leveraging their complementary for an improved performance.
- We augment the input information for our Relation Modules by embedding their architecture with pair-wise mechanisms to explicitly extract relevant extra information from the input pair of joints.
- A new and efficient paradigm for interaction recognition, based on relational networks, as an alternative to the CNN/RNN/GCN architectures that currently dominate the field.

These contributions constitute the novel architecture Interaction Relational Network, which is simple yet effective. We validate on four important datasets that contains human interactions under different conditions, achieving state-of-the-art performance, demonstrating the strength of our technique.

II. RELATED WORK

A. Action Recognition

The problem of human interaction recognition is in fact a sub-field from the more general task of action recognition, which also takes into consideration actions performed by a single person. Previous work with methods using solely RGB videos usually rely on spatio-temporal local features description [30] or CNN architectures for processing appearance and motion separately [31]–[33]. Here we will focus on the previous work that also uses explicit pose information, and have as well evaluated their methods on the SBU interaction dataset.

Most of the latest works using pose information have moved to solutions involving Deep Learning, specially LSTMs [18]–[20], [23]. Du et al. [18] designed a hierarchical structure using

Bidirectional Recurrent Neural Networks (BRNNs), which initially process the joints separately by body part, then fusing these parts step-by-step at subsequent layers, until the whole body is covered. Zhu et al. [19] introduced a regularization in their LSTM to learn the co-occurrence features from the joints – i.e. learn the discriminative joint connections related to human parts. Liu et al. [20] proposed a spatio-temporal LSTM architecture, where the network would unroll not only in the temporal domain, with sequential frames, but also in the spatial, with a pre-defined traversal order of the joints following the human skeleton structure.

Since these works focus on general actions, which mostly are performed by a single person, they do not include in their solutions any strategy specifically to model the relationship between the interacting persons when it is a mutual action. Therefore they disregard the additional information coming from the interaction itself, which our experiments demonstrate to be invaluable when dealing with two-person action recognition.

B. Human Interaction Recognition

Most of early work on human interaction recognition is based on hand-crafted local features, following a bag-of-words paradigm, and often using SVM for the final classification [1], [2], [5], [7], [8]. More recent and successful approaches rely on Convolutional Neural Networks for feature extraction, usually combining it with Recurrent Neural Networks (e.g., Long-short Term Memory) for more complex temporal reasoning over the interactions [3], [9], [10], [34].

Some of these approaches allegedly focus on pose [2], [5], or try to embed human structural knowledge on their solution [34]. Vahdat et al. [5] requires pre-computed person trajectories for their solution, initially they employ a pedestrian detector to obtain bounding boxes around the individuals on the first frame, then subsequently apply a tracker to find the trajectory in the later frames. These bounding boxes are used to retrieve the input information for their solution: spatial location, time of the pose and the cropped image within this region. On the work of Rapitis et al. [2], the authors have labeled pre-defined poselets on some of the frames. They create templates from these poselets by extracting local-features on the cropped image around these annotations. Based on these poselets features, their max-margin framework would then select key-frames to model an action.

Another branch of works do use explicit pose information with the purpose of interaction recognition [6], [13], [14], [16]. For example Yun et al. [6] proposed features coming from different geometric relations between the joints intra-person and inter-person, as well as intra-frames and inter-frames, then using these features as input to an SVM for classification. Ji et al. [13] grouped joints that belonged to the same body part to create poselets, for describing the interaction of these body parts from both individuals, subsequently generating a dictionary representation from these poselets, which will then be fed to an SVM.

Differently from the above mentioned approaches, our proposed architecture does not need the prior knowledge about

the skeleton structure. We do not use the edge information in our architecture and the input is invariant to the joint order, therefore it learns the existing and important relationships by itself.

C. Relational Network

Relational networks have been recently introduced by Santoro et al. [27], targeting to improve relational reasoning over a different range of tasks. It consists of initially reducing the problem input to individual objects, then using a simple architecture that models how these objects relate to each other in a pair-wise form and subsequently uses the general view from all the inferred relationships to solve the problem at hand. The authors demonstrated the versatility of these networks by applying their method on three tasks: visual question answering (QA), text-based question answering and complex reasoning about dynamic physical systems. Which covered not only distinct purposes (question-answering and physical modeling) but also different types of input data (visual, textual and spatial state-based).

Expanding RNs to Video QA, Chowdhury et al. [35] proposed Hierarchical Relational Attention. On which the authors connected attention modules over VGG and C3D features to hierarchical relational modules, nested by input question query token. Park and Kwak [36] applied RNs for 3D Pose Estimation. In their work they used as input estimated 2D joints coordinates, that are grouped by body part and fed in a specific order to their proposed RN to generate the estimated 3D position. Ibrahim et al. [37] also employed Hierarchical RNs, but this time for group activity recognition. They start from CNN features from each individual, then input these features to hierarchically stacked multiple layers of relational modules, each one with their own parameters and relationships graph, which dictates the pairwise relationships to be made.

As far as we know, none of the previous work have designed relational networks that use pose information for the purpose of Human Interaction Recognition, with our work being the first one to extend RNs for this domain and application.

III. RELATIONAL NETWORK OVERVIEW

The Relational Network proposed by Santoro et al. [27] can be simplified through the following equation:

$$RN(O) = f_\phi \left(\sum_{i,k} g_\theta \left(o_i, o_k \right) \right) \qquad (1)$$

Where O is a set of objects $\{o_1, o_2, ..., o_N\}$, on which any i^{th} object is represented by an arbitrary $\mathbb{R}^m$ vector containing the properties of that object. The function g, with learnable parameters θ, is responsible by modeling the relationships for each pair of input objects independently, therefore being also referred to as Relational Module. Meanwhile function f, with trainable parameters ϕ, is in charge for reasoning from the merger of the relationships inferred by g_θ. It is a versatile formulation which, as long as the objects are fed in pairs to the relational module, can have as input different types of data
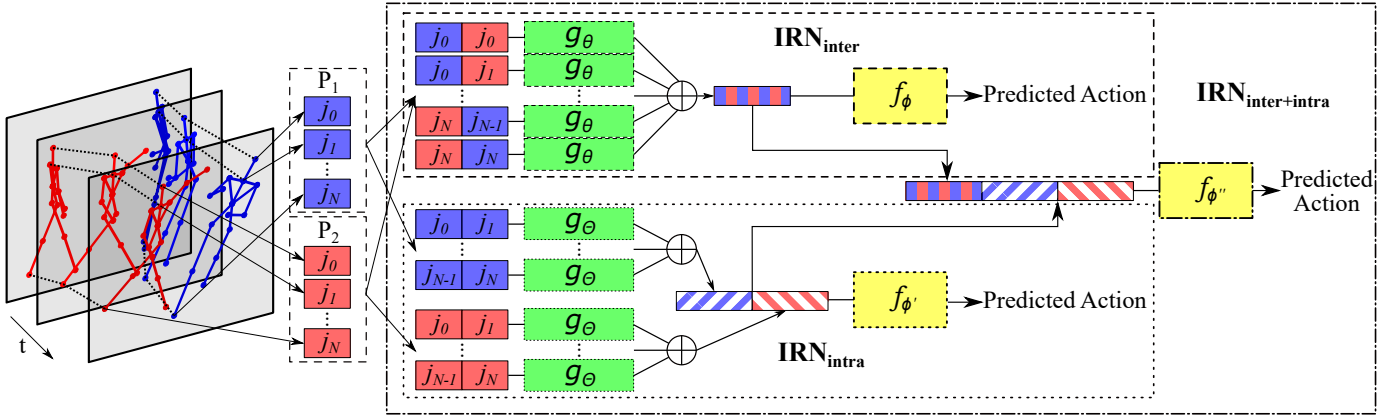
Fig. 2. Illustration of the IRN architecture for human interaction recognition. First, we extract the information across frames from each joint separately (j_n). Then we use the set of joints from both persons (P_p) as input to our different architectures, IRN_{inter} and IRN_{intra}. Each architecture models different relationships between the joints, and can be used independently to predict the action. Furthermore, the models can be fused as $IRN_{inter+intra}$, this way using the knowledge from both types of relationships for a more accurate prediction.

describing the objects, such as CNN/LSTM features or even physical states.

However, the original RN does not cover the case when well-distinguished relationships can be defined due to the nature of the objects and the chosen pairing up rule, leading to distinct specialized models for each type. Missing therefore good ways to train these models and properly fusing them afterward. Also, in its base formulation the Relational Network only contains an unstructured function for reasoning over the input pair, represented by g_θ, lacking a pair-wise structured function that can guide the general relational module with important domain-knowledge information. In the process of tailoring the RN paradigm to our domain, we design important novel extensions to address these issues on our proposed Interaction Relational Network, detailed next.

IV. INTERACTION RELATIONAL NETWORK

We designed an architecture, namely Interaction Relational Network, tailored for human interaction recognition using pose information, inspired by the relational network paradigm of reasoning over pair-wise relationships. Fig. 2 contains a visual representation of the proposed method. First we define how to logically re-arrange the skeleton information so it can be used as input to an RN based architecture, then we formulate important relationships mapping as to comprise all relevant relations pertaining to our problem. In order to fully capture the range of relationships types present in the human interaction problem, it was necessary to extend the RN proposal to handle multiple relations models and to fuse them properly. Also, we further extend the architecture with domain-knowledge pair-wise operations over the input at each relation module, allowing the network to explicitly extract information known to be valuable when dealing with pose data: distance and motion. Finally, we couple our architecture with an LSTM, allowing our method to reason over the interactions during the whole video sequence.

A. Skeleton Joints as Independent Objects

The power of the relational networks lies on breaking down the problem into separate interacting objects, and learning specific parameters for describing how each object relate to each other, before merging this information for the classification of the whole. In the case of pose information and action recognition, we can define each joint j_i as an object, using as their low-level feature its coordinates along the frames, together with integers to identify to which joint and body part does it belong: $j_i = (x_1, y_1, x_2, y_2, ..., x_T, y_T, i, b)$, where x_t and y_t are the 2D coordinates of the joint i belonging to the body part b at the frame t, and T is the desired sampling of frames to be used. As previous work have done [18], [23], [28], we considered five body parts: torso, left hand, right hand, left leg, and right leg. Each person p will therefore have a set of joints for each video, which can be defined as: $P_p = \{j_1^p, j_2^p, ..., j_N^p\}$, where N is the total number of joints provided by the pose data.

B. Relationships Architectures

The simplest approach would be to mix and match all the joints, learning a single model with all pairs combinations. However, besides being computationally inefficient, this naive strategy disregards the fact that the joints might come from other individuals, hence the relationship which each other carries different meaning than when the joints are from the same individual. To allow our method to distinguish between the relevant relationships, and therefore better represent human interactions, we propose exclusive relationships mappings, which will be later merged into a single architecture.

Inter-person Relationships. Since we are dealing with interaction recognition, it is desired to map the relation between the joints inter-person. Therefore, all joints from one individual will be paired with all joints from the other individual, bi-directionally, so that it can be order-invariant (in the case of active/passive actions such as kicking). For that purpose, inspired by the formulation (1), we derive the following equation:

$$IRN_{inter}(P_1, P_2) = f_\phi \left(\sum_{i,k} g_\theta \left(j_i^1, j_k^2\right) \oplus \sum_{i,k} g_\theta \left(j_i^2, j_k^1\right) \right)$$
(2)

Where f_ϕ and g_θ can be Multi-Layer Perceptrons (MLPs), with learnable parameters ϕ and θ respectively. In theory $\sum$ and $\oplus$ can be any pooling operation, such as sum, max, average or concatenate, but from our experiments, we decided to use average because it gives the best results.

Intra-person Relationships. Since the intra-person relationship of the joints can be highly informative as well, we also propose another architecture, where the joints from each person will be paired with the other joints from the same person. For this case there is no need to pair bi-directionally, since the paired joints are from the same person, what would only add unnecessary redundancy to our model – in fact our preliminary experiments demonstrated that it can lead to overfitting in some cases. The pooled output from each individual is concatenated ($\frown$) before going through function f with trainable parameters ϕ'.

$$IRN_{intra}(P_1, P_2) =$$
$$f_{\phi'} \left(\sum_{i=1}^{N} \sum_{k=i+1}^{N} g_\Theta \left(j_i^1, j_k^1\right) \frown \sum_{i=1}^{N} \sum_{k=i+1}^{N} g_\Theta \left(j_i^2, j_k^2\right) \right)$$
(3)

Fusing Relationships. Conclusively, we propose an architecture that fuses both types of relationships under the same function f (parameters ϕ''), by concatenating the pooled information from each function g, each with its own parameters θ and Θ:

$$IRN_{inter+intra}(P_1, P_2) =$$
$$f_{\phi''} \left(\sum_{i,k} g_\theta \left(j_i^1, j_k^2\right) \oplus \sum_{i,k} g_\theta \left(j_i^2, j_k^1\right) \frown \right.$$
$$\left. \sum_{i=1}^{N} \sum_{k=i+1}^{N} g_\Theta \left(j_i^1, j_k^1\right) \frown \sum_{i=1}^{N} \sum_{k=i+1}^{N} g_\Theta \left(j_i^2, j_k^2\right) \right)$$
(4)

Alternatively, we also designed another fusion architecture: $IRN_{inter+intra}^{fc1}$ (Fig. 3). At this design we concatenate the output from f first convolutional layer ($fc1$), of $inter$ and $intra$ models (ϕ and ϕ' parameters respectively), before feeding into the fusion $f_{\phi''}$. The motivation behind this design is to allow us to transfer knowledge not only from the relational modules g_θ and g_Θ, but also from their exclusive f modules. The modules f_ϕ and $f_{\phi'}$ should be specialized on the pooled descriptor for its respective type of relationships, hence contributing with a higher level description for the fusion.

C. Pair-wise Structured Input

One of the advantages of RNs highlighted by Santoro et al. [27] is the flexibility it has to reason over relatively unstructured inputs, for example embeddings from CNNs and LSTMs, as well as it can handle raw data such as state
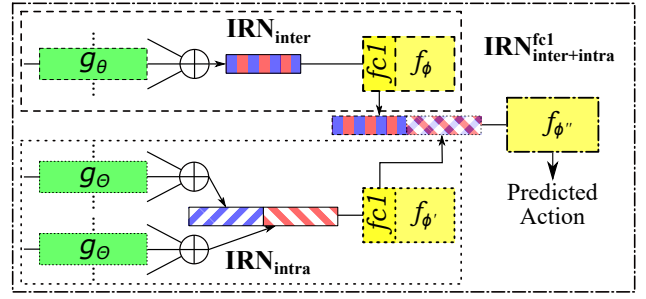

Fig. 3. Simplified view of $IRN_{inter+intra}^{fc1}$, an alternative way from fusing both types of relationships. On this case, the fusion occurs after the first convolutional layer ($fc1$) from f_ϕ and $f_{\phi'}$, which provides a higher level description for the fusion $f_{\phi''}$.

descriptors and, in our case, joints coordinates. Nonetheless, when dealing with structured data it is advantageous to design the architecture with mechanisms to leverage the intrinsic information contained in the data structure. Although this is the case with CNNs and LSTMs used for extracting the embeddings used as input to the RN, they are applied considering only the structure per-object independently. In the case of joints coordinates, the input pair itself is organized in a manner that there is a well-defined structure among them, i.e. coordinates at the same location in the array represents the spatial location of the objects at the same particular time in the video sequence. Therefore it can also be designed architectures to take leverage of the pair-wise structure of the input.

In order to enable the RN to allow this desired feature, but also keeping its flexibility, we propose the following extension to the base formulation (1), on which function h can be carefully designed to model important domain-knowledge pair-wise information:

$$RN(O) = f_\phi \left(\sum_{i,k} g_\theta \left(o_i, o_k, h\left(o_i, o_k\right)\right) \right)$$
(5)

Inspired by previous work that demonstrated the usefulness of explicit distance and motion features for pose-based action recognition [6], [16], we decided to augment our relational network with modules that can automatically extract this information from each pair of joints. In other words, our h function, to be also used as extra input by g_θ, will explicitly compute vectors of distances ($D(j_i, j_k)$) and motions ($M(j_i, j_k)$) from the input object-pair:

$$h\left(j_i, j_k\right) = \left(D(j_i, j_k) \frown M(j_i, j_k)\right)$$
(6)

Considering c_t^i the coordinates of joint i at the frame t, the vector of distances is basically the euclidean distance between the input joints at each frame.

$$D(j_i, j_k) = \left(\left\|c_1^i - c_1^k\right\|, \left\|c_2^i - c_2^k\right\|, ..., \left\|c_T^i - c_T^k\right\|\right)$$
(7)

And the motion vector is defined as the distance between the joints, but at sequential frames:

$$M(j_i, j_k) = \left(\left\|c_1^i - c_2^k\right\|, \left\|c_2^i - c_3^k\right\|, ..., \left\|c_{T-1}^i - c_T^k\right\|\right)$$
(8)

So we can write the improved g_θ, used by IRN', as:

$$g'_\theta(j_i, j_k) = g_\theta(j_i, j_k, h(j_i, j_k)) \qquad (9)$$

D. Sequential Relational Reasoning

Until this point, our proposed IRN could only reason on the interactions over a short period of time. Defined by the number of frames (T) sampled when assembling the joints information (j_i) for input to our network. In other words, the hyper-parameter T would be analogous to a temporal receptive field of our IRN. Preliminary experiments showed that using a too large value of T is detrimental to performance. It was more effective to skip some of the frames by some step size, as in dilated convolutions [38], in order to increase the temporal range of the input. However, this workaround still does not allow the IRN to use the information available in the whole video sequence, and moreover, it has no means to model the intrinsic order of movements contained in an interaction.

Seeking to enable our method to use information from all frames and reason over the evolution of the interactions over time, we incorporate an LSTM to our architecture. The LSTM module is at the end of our architecture, after f_ϕ and before predicting the action. This architecture, hereinafter referenced as LSTM-IRN, can then leverage the information from the whole interaction sequence before classifying the video. Resulting in an architecture also suitable to model over time the relationships present in human interactions.

V. EXPERIMENTS

A. Datasets

SBU [6] is a dataset for two-person interaction recognition, created using Kinect, providing reliable RGBD data for each video. It consists of eight interactions (approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands), seven different participants (pairing up to 21 different permutations), on a total of 282 short videos (around 2-3 seconds each). Recording was done in a single laboratory environment and with a frame rate of 15 frames per second (FPS). Pose information is provided by means of 3D coordinates over 15 joints per person, at each frame. Coordinates are not entirely accurate, containing noise and incorrect tracking at some cases. We followed the 5-fold cross validation protocol defined by the authors, reporting the average accuracy.

UT-Interaction [1] is also a dataset that focus on inter-person interaction, but differently from SBU, it only contains RGB information – i.e. no explicit pose information is provided. To overcome this issue we used OpenPose [12] to extract this type of information, estimating the joints coordinates for the actors in the video. This dataset contains six different classes of actions (shake-hands, point, hug, push, kick, and punch). It was recorded under two different set of conditions, therefore it is subdivided into UT-1 and UT-2, comprising a total of 120 videos, half at each set. Set 1 was recorded on a parking lot, with static background and little camera jitter, meanwhile, Set 2 was taken on a lawn, with partial occlusion from tree branches and with interference from the wind (camera jitter and moving leaves). Videos have about 2-6 seconds each, and a frame rate of 30 FPS. Authors protocol for evaluation consists of a 10-fold cross validation per set.

NTU RGB + D [28] is a dataset with a wide range of general actions. It is not a dataset exclusively for inter-action recognition, however it contains 11 classes of mutual actions (punch/slap, pat on the back, giving something, walking towards, kicking, point finger, touch pocket, walking apart, pushing, hugging, handshaking), more than SBU and UT-Interaction. Moreover, this subset with interaction-only classes, contains a total of 10,347 videos, which were collected with the more precise Kinect (V2) and under more challenging conditions, with 40 different subjects and large variation in viewpoints, by using three cameras recording at the same time. The length of the videos range from 1 to 7 seconds, with a frame rate of 30 FPS. The dataset contains the 3D coordinates from 25 joints per person for all frames. For evaluation, the authors proposed two protocols: Cross Subject (CS), on which 20 pre-defined actors are used for training, and the remaining for testing; and Cross View (CV), where two-cameras are reserved for training, and the third one for testing. Although this is not a dataset specifically for Human Interaction Recognition, we believe that experimenting over this dataset mutual-only classes can be highly valuable at validating our methods because of its characteristics: large scale and more challenging conditions.

NTU RGB + D 120 [29] is an extension from the dataset above, and it is the largest available dataset for human action recognition with RGBD data. It contains 60 additional classes, of which 15 are mutual-action classes (hit with object, wield knife, knock over, grab stuff, shoot with gun, step on foot, high-five, cheers and drink, carry object, take a photo, follow, whisper, exchange things, support somebody, rock-paper-scissors). In total there are 24,794 videos from 26 interaction-only classes, captured similarly as the previous version: using Kinect (V2), with many different subjects and a diverse range of viewpoints. The evaluation protocol is slightly different: Cross-Subject, which is the same as before, with a subset of actors being used exclusively for training and another for testing; and Cross-Setup, where instead of splitting the data according to the camera, pre-defined setups are selected for comprising the training and testing splits.

B. Implementation Details

MLPs Configuration. Hyperparameters detailed here were tuned during preliminary experiments. The IRN is implemented as an MLP, where g_θ consists of four fully-connected layers, the first three with 1000 units and the last with 500, and f_ϕ contains three fully-connected layers with 500, 250 and 250 units respectively, with dropout rate of 0.25. The LSTM module contains 256 units and also a dropout of 0.25. A softmax layer is placed at the end of the architecture to generate the interactions prediction scores. Training was performed with Adam optimizer, learning rate value of 1e-4 and weight initialization following a truncated normal distribution with zero mean and 0.045 standard deviation for SBU and UT, and 0.09 standard deviation for NTU RGB+D and NTU RGB + D 120.
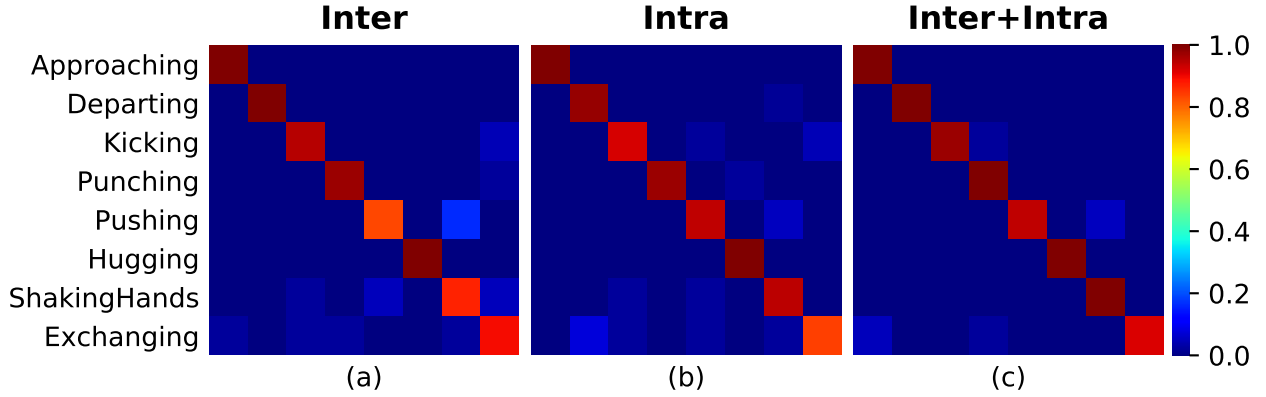
Fig. 4. Confusion matrices for SBU dataset with methods: (a) LSTM-IRN'_{inter}, (b) LSTM-IRN'_{intra} and (c) LSTM-$IRN'^{fc1}_{inter+intra}$.

Training Procedure. To improve generalization during training, we randomly swap the input order between the persons' joints set ($P_1 \rightleftharpoons P_2$). This was significantly beneficial for the IRN_{intra} architecture to avoid bias on the order of the concatenated feature generated after g_Θ. $IRN_{inter+intra}$ parameters θ and Θ are fine-tuned from the weights obtained previously by training IRN_{inter} and IRN_{intra} separately, meanwhile ϕ is randomly initialized. For $IRN^{fc1}_{inter+intra}$, in addition to θ and Θ, we also use the $fc1$ layer and weights from previously trained IRN_{inter} and IRN_{intra}, for initializing the model before training.

Pose Estimation. OpenPose [12] is a readily available toolbox for extracting pose information from RGB videos. We ran it with its default options, which were not the most accurate, but it was faster and sufficiently precise for our experiments. However, because the output is frame-based, we had to apply some post-processing steps to guarantee the consistency of the bodies from Person 1 and 2 throughout the video. It consisted mainly on correctly assigning the pose at each frame to its respective body, based on the distance of the joints between the current frame and previous frames, assuming the overall distance should not be too different between consecutive frames. This is important when there are more than two poses in the frame, due to noise or because of passersby, or when the order of the actors' pose change.

Joints and Frames Subsampling. Although the poses for UT (estimated) and NTU RGB+D (provided) contains 25 joints, we sampled only 15 of them, analogous to what is provided by SBU data. As for the parameter T, regarding the temporal receptive field, we use different values for each dataset. For SBU, since the videos are shorter, we use 8 consecutive frames as a sampling for our input feature. Since UT, NTU RGB+D and NTU RGB+D 120 have longer videos, we use 32 frames. For the experiments without an LSTM we have chosen to sample the central frames, because most likely they contain the more relevant parts of the interaction. For LSTM-IRN we build our input sequence with overlapping frames at each timestep, by sampling T frames of the video with a step of $T/2$ frames, i.e. 8 frames input for every 4 frames step for SBU and 32 frames every 16 for the other datasets.

C. Ablation experiments

First, to better evaluate the impact of each part from our proposed methodology, we separately show our results on the SBU dataset, reported in Table I.

TABLE I
ABLATION EXPERIMENTS ON SBU DATASET.

Experiment	Acc
IRN_{inter} (Baseline)	88.7%
IRN'_{inter} (Augmented Input)	93.6%
LSTM-IRN'_{inter}	94.6%
IRN_{intra} (Baseline)	95.4%
IRN'_{intra} (Augmented Input)	95.8%
LSTM-IRN'_{intra}	95.2%
Naive-$IRN'_{inter+intra}$	92.0%
Averaging scores	94.3%
Random-$IRN'_{inter+intra}$	92.5%
$IRN'_{inter+intra}$	96.1%
$IRN'^{fc1}_{inter+intra}$	96.7%
LSTM-$IRN'^{fc1}_{inter+intra}$	98.2%

Our baseline, using only the central frames coordinates and joint indexing information, already obtains 88.7% of accuracy with IRN_{inter} and 95.4% of accuracy with IRN_{intra} architecture. These results indicates that our approach is able to successfully map the different types of relationships present in the problem of interaction recognition. Incorporating our IRN with mechanisms to take leverage of the pair-wise structure of the input, what allows the network to extract explicit distance and motion information, proved to be highly valuable to IRN_{inter} and it was also helpful to IRN_{intra}, increasing the performance on both cases.

The Naive-$IRN'_{inter+intra}$ experiment consists on simply merging all the relationships mapping ($inter$ and $intra$) into a single model relational network. This means only one set of parameters for g and pooling the outputs indiscriminately into the same global descriptor before f. The Naive-$IRN'_{inter+intra}$ approach underperforms, obtain-

ing only 92.0% of accuracy, lower than the specialized IRN' models. Averaging the scores from these specialized models is better than naively merging the relationships, but it is still worse than using IRN'_{intra} alone.

We were able to truly take leverage of the complementary between the relationships models only through our proposed fused architecture $IRN'_{inter+intra}$. This method maintains the specialized models not only on its design, but also by initially training them separately – random initialization of weights (Random-$IRN'_{inter+intra}$) also underperforms. Furthermore, our $IRN'^{fc1}_{inter+intra}$ approach seems to further correlate and benefit from the complementary of the relationships, obtaining a more significant improvement. Ultimately, by coupling our IRN with the LSTM module we extend the relational reasoning to all available frames and enable reasoning over the interaction sequence, obtaining our best result: 98.2% with LSTM-$IRN'^{fc1}_{inter+intra}$.

Fig. 4 contains the confusion matrices for all LSTM-IRN architectures. It is interesting to notice how the two relationships models have confusion on different interactions, for example $Inter$ have some confusion between $Punching$ and $Exchanging$ while $Intra$ does not, and how this confusion is greatly reduced at $Inter+Intra$. More importantly, interaction classes getting confused by both models, such as $Pushing$ and $ShakingHands$, are almost entirely distinguished with $Inter+Intra$. This qualitative analysis is a good indication on the capability of our proposed architecture to keep the strength of both types of relationships models, and also to leverage their complementary for distinguishing between even harder cases.

D. State-of-the-art Comparisons

TABLE II
COMPARISON OF OUR RESULTS WITH PREVIOUS WORK ON SBU.

Method	Acc
Yun et al. [6]	80.3%
Ji et al. [13]	86.9%
HBRNN [18] (reported by [19])	80.4%
CHARM [39]	83.9%
CFDM [14]	89.4%
Co-occurrence LSTM [19]	90.4%
Deep LSTM (reported by [19])	86.0%
ST-LSTM [20]	93.3%
VA-LSTM [40]	97.2%
Wu et al. [16]	91.0%
Two-stream GCA-LSTM [23]	94.9%
LSTM-IRN'_{inter}	94.6%
LSTM-IRN'_{intra}	95.2%
LSTM-$IRN'^{fc1}_{inter+intra}$	98.2%

SBU. We report our best results for SBU dataset in Table II, alongside previous work results. Our specialized relationship architectures, LSTM-IRN'_{inter} and LSTM-IRN'_{intra}, can already outperform or be comparable to almost

all of previous work. After fusion through architecture LSTM-$IRN'^{fc1}_{inter+intra}$ we obtain state-of-the-art performance, with an advantage over 1 percentage point over the second best: 97.2% with VA-LSTM [40].

TABLE III
RESULTS FROM PREVIOUS APPROACHES AND OUR PROPOSED METHODS ON THE UT DATASET PER SET.

Method	Acc	
	UT-1	UT-2
Ryoo et al. [1]	70.8%	-
Raptis et al. [2]	93.3%	-
Donahue et al. [9]	85.0%	-
Kong and Fu [41]	93.3%	91.7%
Wang and Ji [4]	95.0%	-
Aliakbarian et al. [3]	90.0%	-
Ke et al. [34]	93.3%	91.7%
Shi et al. [10]	97.0%	-
Shu et al. [42]	98.3%	-
LSTM-IRN'_{inter}	93.3%	96.7%
LSTM-IRN'_{intra}	96.7%	91.7%
LSTM-$IRN'^{fc1}_{inter+intra}$	98.3%	96.7%

UT-Interaction. Table III shows our results on the UT dataset, together with the results from previous works. The LSTM-$IRN'^{fc1}_{inter+intra}$ obtains a performance equivalent to the current state-of-the-art on the set UT-1, an almost saturated accuracy of 98.3%. On UT-2 set, our architectures outperform the previous methods, and set the new state-of-the-art to 96.7%, which is also very high. For UT-2 there was no improvement from using LSTM-$IRN'^{fc1}_{inter+intra}$, when in comparison to the specialized LSTM-IRN'_{inter} model, it might be the case that for this dataset the interactions are not that complex, therefore not benefiting so much from more advanced relationships modeling. To the best of our knowledge, our work is the first using explicit pose information at UT-Interaction dataset, with the previous works being based on RGB information only. Therefore, our experiments demonstrate that it is feasible and valuable to extract the pose from RGB videos before performing human interaction recognition.

NTU RGB+D. Our experiments results over the interaction classes of this dataset are present in Table IV. Both of the specialized architectures are able to match or outperform the compared methods on the two protocols. LSTM-IRN'_{inter} outperforms with a significantly superior performance, demonstrating the importance of considering the inter-person joints relationships when dealing with mutual actions. Our LSTM-$IRN'^{fc1}_{inter+intra}$ architecture can obtain an even higher performance, going beyond 90% accuracy not only on the Cross-View protocol, but also on the more challenging Cross-Subject.

NTU RGB+D 120. The new version of the NTU RGB+D dataset [29] is even more challenging, as shown in Table V containing our results together with previous work methods. Our proposed LSTM-IRN architectures still obtains state-of-

TABLE IV
RESULTS FROM OUR PROPOSED METHODS ON THE SUBSET OF NTU RGB+D DATASET, CONTAINING ONLY THE MUTUAL ACTION CLASSES.

Method	Acc Mutual Actions	
	Cross-Subject	Cross-View
ST-LSTM [20]	83.0%	87.3%
GCA-LSTM [43]	85.9%	89.0%
Two-stream GCA-LSTM [23]	87.2%	89.9%
FSNET [44]	74.0%	80.5%
LSTM-IRN'_{inter}	89.5%	92.8%
LSTM-IRN'_{intra}	87.3%	91.7%
LSTM-$IRN'^{fc1}_{inter+intra}$	90.5%	93.5%

TABLE V
RESULTS FROM OUR PROPOSED METHODS ON THE MUTUAL-ACTIONS SUBSET OF NTU RGB+D 120 DATASET.

Method	Acc Mutual Actions	
	Cross-Subject	Cross-Setup
ST-LSTM [20]	63.0%	66.6%
GCA-LSTM [43]	70.6%	73.7%
Two-stream GCA-LSTM [23]	73.0%	73.3%
FSNET [44]	61.2%	69.7%
LSTM-IRN'_{inter}	74.3%	75.6%
LSTM-IRN'_{intra}	73.6%	75.2%
LSTM-$IRN'^{fc1}_{inter+intra}$	77.7%	79.6%

Fig. 5. Performance per interaction class of method LSTM-$IRN'^{fc1}_{inter+intra}$ on the NTU RGB+D 120 dataset, for both protocols.

the-art performance, what indicates they can scale with respect to the number of classes, so they are also suitable at describing and distinguishing between a larger number of challenging interactions. Moreover, as the complexity of the interaction recognition problem grows, more benefit can be obtained with the multi-relationship model LSTM-$IRN'^{fc1}_{inter+intra}$. Claim supported by the fact that at NTU RGB+D the gain in performance obtained was approximately of 1 percentage point over LSTM-IRN'_{inter}, meanwhile at NTU RGB+D 120 this gain was between 3 and 4 percentage points.

A more detailed examination of the performance of LSTM-$IRN'^{fc1}_{inter+intra}$ is shown in Fig. 5, which contains the accuracy per interaction class for both protocols. Some interactions are significantly more challenging than others, with a recognition rate much inferior than the average. Through the confusion matrices present in Fig. 6 it can be seen that, at least for the most severe cases (around or bellow 70% accuracy), both protocols have confusion mostly between the same interactions. A few notable differences are the confusion between *SupportSomebody* and *KnockOver* for Cross-Subject, and *TouchingPocket* with *PattingOnBack* for Cross-Setup. On both protocols there is a lot of confusion between *HitWithObject* and *WieldKnife*, and these two with *Punch/slapping*. Also, there is some confusion between *Give-Something* with *Handshaking* and *ExchangeThings*, as well as
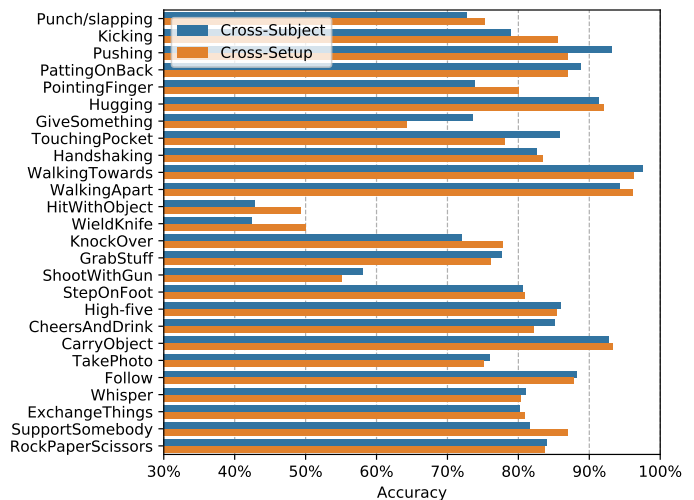
between *ShootWithGun* and *TakePhoto*. These confusions are reasonable, since we are using only the body parts coordinates information, and these interactions should have similar human movements (e.g. extending the hands to each other), being mainly distinguishable by the object (or its absence) each individual holds. We believe this is a strong evidence that for more advanced interaction recognition, RGB visual information should also be used.

VI. CONCLUSION

In this work, we proposed the novel Interaction Relational Network for recognition of mutual actions, through an architecture focusing at relational reasoning over the different relationships between the human joints during interactions. Current techniques dominating the field of pose-based interaction recognition are based on CNNs/LSTMs/GCNs and consist of fixed structures determining how to relate each joint. Therefore, given our IRN capacity to learn from the data itself how the human body parts relate to each other while two individuals interact, it stands as a highly valuable alternative to these techniques.

Our proposed method obtains state-of-the-art performance on the traditional interaction datasets SBU and UT, and it also obtains the highest performance on the mutual actions subset of the datasets NTU RGB+D and NTU RGB+D 120. To accomplish such deed, important extensions had to be made over the original Relational Network proposition in order to tailor it for our problem and data. Extensions such as: multiple relationships models; fusion of these multiple relationship models; domain knowledge pair-wise structured operation over the input; and sequential relational reasoning. We hope these extensions can also be valuable to other problems, with specific data and relationship particularities that the original RN it is not capable to suitably model without it. A more complex scenario that we believe to be a strong candidate for benefiting from our architecture and extensions, after some adaptations, is the problem of Group Activity Recognition [37], [45].
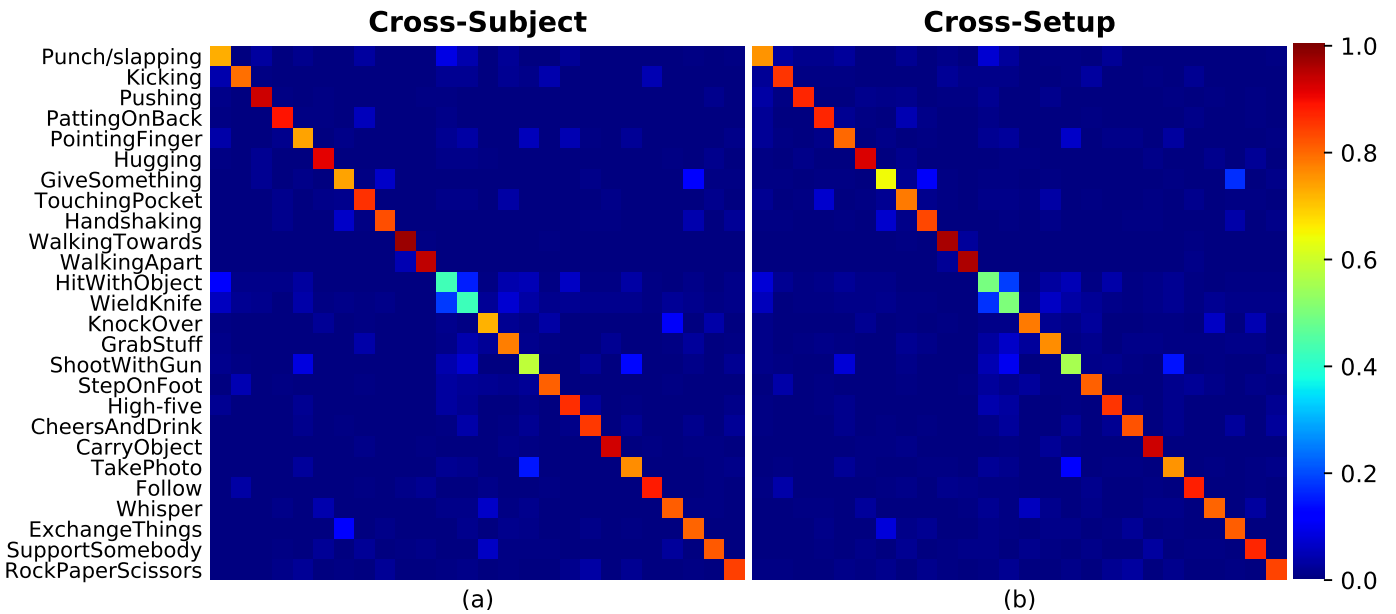
Fig. 6. Confusion matrices for NTU RGB+D 120 dataset with method LSTM-$IRN'^{fc1}_{inter+intra}$ for protocols (a) Cross-Subject and (b) Cross-Setup.

Furthermore, we believe our IRN can still be improved through different means. Our relational module can benefit from other sources of data, such as high-level visual features extracted on the region around the joints coordinates, which can be appended to the current input as extra and valuable information. Another possible improvement, would be to design a more sophisticated pair-wise structured operation. Alternatively to the fixed operations for distance and motion, which are based on the knowledge that the input data represents spatial coordinates at different frames, an architecture with trainable parameters can be designed to take leverage of the same knowledge. A different potential place for improvement in our architecture is at the pooling stage. Instead of averaging all pairs features with the same weight, therefore giving the same importance to all pairs of joints, an attention mechanism can be applied, so as to adjust the contribution to the global descriptor depending of which joints does that relationship belong to.

ACKNOWLEDGMENT

REFERENCES

[1] M. S. Ryoo and J. K. Aggarwal, "Spatio-Temporal Relationship Match : Video Structure Comparison for Recognition of Complex Human Activities," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 1593–1600.

[2] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2650–2657.

[3] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging LSTMs to Anticipate Actions Very Early," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 280–289.

[4] X. Wang and Q. Ji, "Hierarchical context modeling for video event recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 9, pp. 1770–1782, 2017.

[5] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori, "A discriminative key pose sequence model for recognizing human interactions," in *IEEE International Conference on Computer Vision*, 2011, pp. 1729–1736.

[6] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 28–35.

[7] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming video," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1036–1043.

[8] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen, "Spatio-Temporal Phrases for Activity Recognition," in *Springer European Conference on Computer Vision (ECCV)*, 2012, pp. 707–721.

[9] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[10] Y. Shi, B. Fernando, and R. Hartley, "Action Anticipation with RBF Kernelized Feature Mapping RNN," in *Springer European Conference on Computer Vision (ECCV)*, 2018, pp. 305–322.

[11] Z. Zhang, "Microsoft Kinect Sensor and Its Effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *arXiv preprint arXiv:1812.08008*, pp. 1–14, 2018.

[13] Y. Ji, G. Ye, and H. Cheng, "Interactive body part contrast mining for human interaction recognition," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2014, pp. 1–6.

[14] Y. Ji, H. Cheng, Y. Zheng, and H. Li, "Learning contrastive feature distribution model for interaction recognition," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 340–349, 2015.

[15] M. Li and H. Leung, "Multiview Skeletal Interaction Recognition Using Active Joint Interaction Graph," *IEEE Transactions on Multimedia (TMM)*, vol. 18, no. 11, pp. 2293–2302, 2016.

[16] H. Wu, J. Shao, X. Xu, Y. Ji, F. Shen, and H. T. Shen, "Recognition and Detection of Two-Person Interactive Actions Using Automatically Selected Skeleton Features," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 3, pp. 304–310, 2018.

[17] M. Perez, J. Liu, and A. C. Kot, "Interaction Recognition Through Body Parts Relation Reasoning," in *Springer Asian Conference on Pattern Recognition (ACPR)*, 2019, pp. 1–12.

[18] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network

for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[19] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, and Others, "Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks." in *AAAI*, vol. 2, no. 5, 2016, pp. 3697–3703.

[20] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition," in *Springer European Conference on Computer Vision (ECCV)*, 2016, pp. 816–833.

[21] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective Active Skeleton Representation for Low Latency Human Action Recognition," *IEEE Transactions on Multimedia (TMM)*, vol. 18, no. 2, pp. 141–154, 2016.

[22] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative Multi-instance Multitask Learning for 3D Action Recognition," *IEEE Transactions on Multimedia (TMM)*, vol. 19, no. 3, pp. 519–529, 2017.

[23] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 4, pp. 1586–1599, 2018.

[24] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning Clip Representations for Skeleton-Based 3D Action Recognition," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 6, pp. 2842–2855, 2018.

[25] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Learning Latent Global Network for Skeleton-based Action Prediction," *IEEE Transactions on Image Processing (TIP)*, pp. 1–12, 2019.

[26] J. Liu, H. Rahmani, N. Akhtar, and A. Mian, "Learning Human Pose Models from Synthesized Data for Robust RGB-D Action Recognition," *Springer International Journal of Computer Vision (IJCV)*, vol. 127, no. 10, pp. 1545–1564, 2019.

[27] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 4967–4976.

[28] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1010–1019.

[29] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2019.

[30] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3551–3558.

[31] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[32] 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," in *Springer European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.

[33] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential Deep Trajectory Descriptor for Action Recognition With Three-Stream CNN," *IEEE Transactions on Multimedia (TMM)*, vol. 19, no. 7, pp. 1510–1520, 2017.

[34] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Leveraging Structural Context Models and Ranking Score Fusion for Human Interaction Prediction," *IEEE Transactions on Multimedia (TMM)*, vol. 20, no. 7, pp. 1712–1723, 2018.

[35] M. I. H. Chowdhury, K. Nguyen, S. Sridharan, and C. Fookes, "Hierarchical Relational Attention for Video Question Answering," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 599–603.

[36] S. Park and N. Kwak, "3D Human Pose Estimation with Relational Networks," in *British Machine Vision Conference (BMVC)*, 2018, pp. 1–13.

[37] M. S. Ibrahim and G. Mori, "Hierarchical Relational Networks for Group Activity Recognition and Retrieval," in *Springer European Conference on Computer Vision (ECCV)*, 2018, pp. 721–736.

[38] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1–13.

[39] W. Li, L. Wen, M. C. Chuah, and S. Lyu, "Category-Blind Human Action Recognition: A Practical Recognition System," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4444–4452.

[40] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2136–2145.

[41] Y. Kong and Y. Fu, "Close human interaction recognition using patch-aware models," *IEEE Transactions on Image Processing (TIP)*, vol. 25, no. 1, pp. 167–178, 2016.

[42] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition," *CoRR*, vol. 1811.00270, pp. 1–8, 2018.

[43] J. Liu, G. Wang, L.-Y. Duan, P. Hu, and A. C. Kot, "Global Context-Aware Attention LSTM Networks for 3D Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1647–1656.

[44] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Skeleton-Based Online Action Prediction Using Scale Selection Network," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2019.

[45] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1971–1980.

Mauricio Perez is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He received his M.Sc. degree from University of Campinas in 2016, and his B.Sc. degree from Federal University of So Carlos, both in Brazil.

His research interests include Deep Learning, Video Analysis, Sensitive Media Detection, Medical Imaging and Computer Vision.
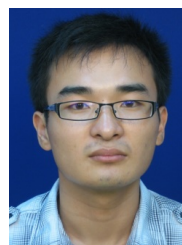
Jun Liu obtained the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2019, the B.Eng. degree from Central South University, China, in 2011, and the M.Sc. degree from Fudan University, China, in 2014.

His research interests include video analysis, human action recognition, pose estimation, and deep learning.

Alex C. Kot (F06) has been with Nanyang Technological University, Singapore, since 1991. He headed the Division of Information Engineering with the School of Electrical and Electronic Engineering for eight years and served as an Associate Chair (Research). He was the Vice Dean (Research) with the School of Electrical and Electronic Engineering and the Associate Dean for the College of Engineering for eight years. He is currently a Professor with the School of Electrical and Electronic Engineering and the Director of the Rapid-Rich Object Search Lab. He has published extensively in the areas of signal processing for communication, biometrics, image forensics, information security, and computer vision. He was a recipient of the Best Teacher of the Year Award and co-authored several best paper awards, including for ICPR, WIFS, and IWDW. He was awarded as the IEEE Distinguished Lecturer of the Signal Processing Society.

He served as the Associate Editor for the IEEE Transactions onImage Processing (T-IP), IEEE Transactions on Signal Processing (T-SP), IEEE Transactions on Multimedia (T-MM), IEEE Signal Processing Magazine (SPM), IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT), and IEEE Transactions on Information Forensics and Security (T-IFS). He has served the IEEE Signal Processing Society in various capacities, such as the General Co-Chair at the 2004 IEEE International Conference on Image Processing and the Vice President of the IEEE Signal Processing Society. He is a fellow of the Academy of Engineering, Singapore, a Fellow of IEEE, and a Fellow of IES.