Locating Human Interactions With Discriminatively Trained Deformable Pose+Motion Parts

Coert van Gemeren*, Ronald Poppe† and Remco C. Veltkamp‡
Interaction Technology Group, Department of Information and Computing Sciences, Utrecht University
The Netherlands
Email: *C.J.vanGemeren@uu.nl, †R.W.Poppe@uu.nl, ‡R.C.Veltkamp@uu.nl

***Abstract*—We model dyadic (two-person) interactions by discriminatively training a spatio-temporal deformable part model of fine-grained human interactions. All interactions involve at most two persons. Our models are capable of localizing human interactions in unsegmented videos, marking the interactions of interest in space and time. Our contributions are as follows: First, we create a model that localizes human interactions in space and time. Second, our models use multiple pose and motion features per part. Third, we experiment with different ways of training our models discriminatively. When testing on the target class our models achieve a mean average precision score of 0.86. Cross dataset tests show that our models generalize well to different environments.**

I. Introduction

We are interested in detecting fine-grained interactions between people. Detecting such interactions in videos has a wide range of applications in video search, automated video captioning and in surveillance. Modeling human interactions from videos is motivated by these applications and has gained more research interest in recent years [1]–[4].

Human interaction detection is challenging. It involves several subtasks. First, we need to localize the people involved in an interaction. Second, we need to find out who interacts with whom. Third, for each pair, the start and the end of the interaction needs to be determined. Finally, the identified spatio-temporal region is assigned the label of the most likely interaction class. A large body of work has emerged focusing on these subtasks. Progress has been made in human tracking and action classification [5]. However, solving each subtask independently is unlikely to give the best results. Errors made early on in, for instance, the person detection impact the final classification. To overcome these issues, we introduce an approach that models all parts of the problem simultaneously.

The classification of human interactions benefits significantly from precise information of limb positions and movements [6]. The pose or the movement alone is typically not sufficient to distinguish between similar interactions [7]. For instance, Fig. 1 shows two individuals involved in a hand shake and when passing an object. The poses look similar but their motion is different. However, the movement of passing an object and a fist bump can both be characterized by two right hands moving towards each other. In this case, their poses are different. Therefore, we need to model both the pose and the motion of the body parts that are most representative for the interaction. Pose and motion can be described by Histograms of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF), respectively. In this paper, we model body parts with HOG and HOF descriptors within a deformable parts model (DPM), which spatially structures the parts.

We focus on human interactions that vary subtly and have a moment of contact between the two individuals. Our models can be used to localize specific interactions in both space and time, in unsegmented videos. The output is a set of spatio-temporal areas with an assigned interaction label.

When detection models are trained for each interaction class independently, similarities in feature descriptors will arise between similar classes (see Fig. 1, right side). This negatively affects the discriminative power of the models. In this paper we address this issue through discriminative training of the models.

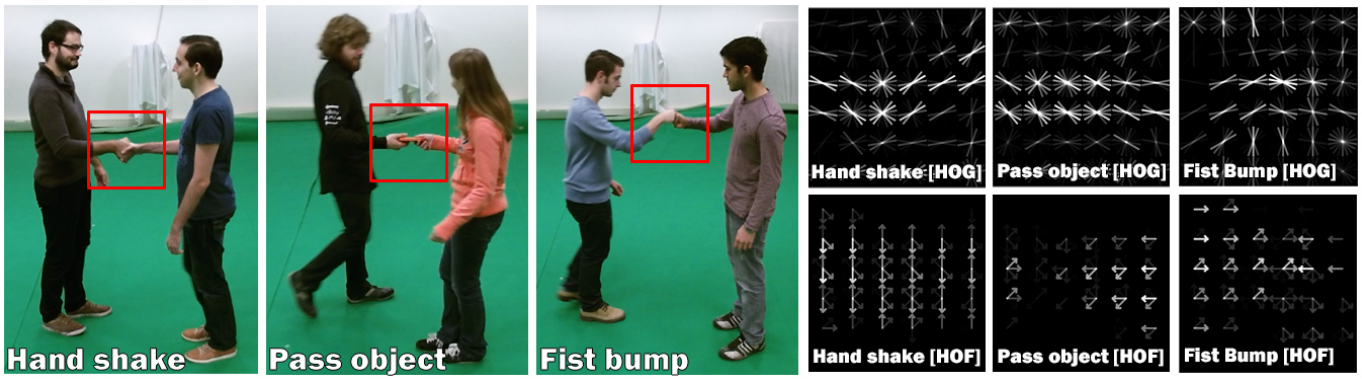


Fig. 1: Three interactions that only have subtle differences in pose (HOG) and movement (HOF). The red bounding boxes show the areas of the right hands of which the part descriptions (HOG and HOF) are shown on the right.

Our contributions are as follows. First, we introduce a framework to localize human interactions in both space and time. Second, we use multiple pose and motion features per part, which enables us to detect fine-grained interactions. Thrid, instead of modeling body parts for each individual, we model specific the most suitable regions of both individuals simultaneously, to represent specific interactions.

We experiment with different ways to train such models discriminatively. We show the efficacy of our work in spatio-temporal localization experiments both on a single dataset, and in a cross-dataset scenario.

Next, we will discuss related work, followed by a detailed explanation of the modeling process in Section III. In Section IV we detail our experiments and discuss their results. We conclude in Section V.

II. Related work

Different strategies can be chosen to solve the problem of interaction detection. One approach is to first find candidate regions of people throughout the video using human detection algorithms [8], [9]. The interactions in these regions can then be classified based on extracted features [9]. Patron-Perez et al. [8] use this two-stage approach to classify human interactions in unsegmented videos. The drawback of this approach is that classification is suboptimal when the person localization fails, for example due to partial occlusions when people are close to each other. The relative distance between individuals has been further explored by Sener and İkizler [10], who formulate interaction detection as a multiple-instance learning problem because not all frames in an interaction are considered informative. Sefidgar et al. [11] use the same reasoning to create a model based on discriminative key frames and consider their relative distance and timing within the interaction.

Another approach is to first generate features without knowing the locations of the people. Features are extracted around keypoints, such as Space-Time Interest Points (STIP) [12]. These features can be encoded using Fisher Vectors (FV) [13] or a Bag-of-Features (BoF) [1]. This approach has achieved state-of-the-art results. Wang et al. [14] create a BoF dictionary based on dense trajectories of keypoints. These trajectories consist of many tracks of features: Histograms of Oriented Gradients (HOG), Histograms of Optical Flow (HOF) and Motion Boundary Histograms (MBH). One drawback of this method is that it is not possible to localize an interaction in space and time because one trajectory by itself is not sufficiently informative to give a definite demarcation of the interaction subspace. Ni et al. [15] solve this problem by clustering the dense trajectories. When enough dense trajectories can be clustered, the volume created by the set of trajectories roughly encompasses the interaction. While this is a promising direction, these methods do not link motion descriptors to specific body parts, As such, they cannot be used to distinguish between interactions that vary subtly.

Poses and movements of specific body parts characterize the interaction that is taking place [16]. For instance, in a fist bump, the arms extend towards each other and the knuckles of the right hands meet as they touch. The clenched fist required for the knuckles to meet means that the lower arm is in a particular pose and moves forward. The notion of motion and pose units has been used by Kong et al. [17], who create models for attributes such as "outstretched hands" and "leaning forward torso" and consider their co-occurrences. Deformable part models (DPM, [18]) also link pose and motion to specific parts of the body. DPMs model spatial structures at a coarse level and then put a number of finer-grained parts on top of the most dominant features. Yao et al. [19] use DPMs to model poses in human-object interactions. To capture the movement related to a given pose, they connect the output of a DPM to a set of motion templates. These templates do not have deformation parameters, which means that they cannot model the variation in movements. Tian et al. [20] have extended DPMs for action detection using HOG3D parts [21] combined with spatio-temporal deformation parameters. In their formulation, each HOG3D part is deformable in both space and time, but is not connected to any specific body parts.

In contrast, Van Gemeren et al. [7] use interaction-specific part based models with HOG descriptors for specific body parts. The coordinated movement between the people is modeled as a HOF descriptor, locally related to the detected persons. As there can be significant variation in how people pose, this two-stage approach strongly relies on the accuracy of the pose detection. In this paper, we address these issues with deformable part models that take into account both pose and motion features simultaneously.

III. Discriminative interaction model

The starting point for our two-person interaction model is the model introduced by Yang and Ramanan [22]. We change three key properties to make it suitable for human interaction detection. First, parts do not model generic body parts at joint locations. Instead, in this work parts model those regions of the body that are most suited to represent specific interactions. This representation is remniscent of the poselets introduced by Bourdev et al. [23]. Second, each part in our model can represent different combinations of image cues. It can represent either gradient or motion features, or it can represent a combination of the two. This way we can explicitly decide per part if it should model pose, motion or both. Finally, we model the spatial relation between the deformable body parts of both persons involved in the interaction simultaneously. In our model variations in the part positions, which implicitly represents the distance between the two persons during the interaction, are optimized as latent deformation parameters of the DPM. These extensions of the model allow us to do spatio-temporal interaction localization for interactions with subtle differences.

A. Model formulation

We base our model on the observation that for fine-grained human interactions there exists a moment where both the pose and motion are coordinated in a way that is prototypical for

the interaction. To model this prototypical moment we define a graph $G = (V, E)$, with V a set of K body parts and E the set of connections between pairs of parts [22]. In the experiments presented in this paper, the body parts we consider may be compound parts consisting of multiple body joints, such as a *torso*, *right upper arm*, *right lower arm* and *right hand*. Each body part i is centered on location $l_i = (x_i, y_i)$. The scoring for a part configuration in image I is given by:

$$S(I, l) = \sum_{i \in P} w_i \cdot \phi_i(I, l_i) + \sum_{ij \in E} w_{ij} \cdot \psi(l_i - l_j) \quad (1)$$

We note that in Eq. 1, for clarity, we omit that the scores are defined by the dot product between a part and a subwindow of a feature pyramid computed from the input image. The first term models the part appearance with a convolution of image feature vector $\phi_i(I, l_i)$ with trained detector w_i. The second term contains the pair-wise deformations between parts $\psi(l_i - l_j) = \left[dx\ dx^2\ dy\ dy^2\right]$, with $dx = r_i x_i - r_j x_j$ and $dy = r_i y_i - r_j y_j$ the relative location of part i with respect to part j [22]. The distances dx and dy are defined with respect to a root factor r. This allows each part to have its own spatial resolution. In practice we use two resolutions, the second of which is twice the resolution of the first. This allows us to quickly find candidate detections at the coarse level, on which a localization of fine-grained pose and motion cues can be detected. This structure makes the model suitable for cascade object detection [24]. w_{ij} encodes the rest location and the rigidity of the connections between parts.

The three key adaptations to this model are defined as follows:

Class-specific part detectors Instead of having different part mixtures representing various orientations, we learn class-specific detectors that encode the articulation of the body parts directly, such as a bent arm or a side-facing torso. Therefore, we use only a single detector per class, instead of a mixture of part detectors like in [22].

Multiple features Our model supports different types of features per part. For part i with feature representations D_i, we replace the first term in Eq. 1 by:

$$\sum_{i \in P} \sum_{j \in D_i} b_{ij} w_i^j \cdot \phi_i^j(I, l_i) \quad (2)$$

$\phi_i^j(I, l_i)$ denotes a feature vector of type j for part i. Bias b_{ij} denotes the weight for each feature type. w_i^j is the trained detector for part i and feature type j. Parts can have different combinations of features D_i. In this work D_i is either HOG or HOF, but it is not limited to these types of features. The DPM inference algorithm is well suited to incorporate a learned feature extractor such as convolutional neural networks (CNN) [25]. As such, our formulation is different from Yao et al. [19], who require one HOG template and a set of HOF templates per body part. In contrast, our model allows us to focus on those features that are characteristic for a specific body part and interaction class. We explicitly also consider features that are calculated over time such as HOF descriptors.

Two-person interaction As there are two persons involved in a dyadic interaction, we combine their body parts into the same graph. Each actor's body parts form a sub-tree in this $(2K + 1)$-node graph. The torso parts of both actors are connected through a virtual root part of the graph. This part does not have an associated part detector but it allows us to model relative distances between people, similar to Patron-Perez et al. [8] and Sener and İkizler [10].

B. Training

For each interaction class, we learn the model from a set of training sequences. We describe a sequence of length n as $X = \{(I_i, y_i, p_i)\}_{i=1}^n$ with I_i an image frame, y_i the interaction label of frame i and p_i a pose vector containing the 2D joint positions of the two persons performing the interaction. We assume the sequences are segmented in time to contain the interaction of interest.

We train a model in three steps. First, we determine the *epitome* frame per training sequence. Second, we learn the initial body part detectors. Third, we simultaneously update the epitome frame and the body part detectors.

Epitome frame detection We intend to find the prototypical interaction frame of each training sequence based on the pose of the individuals. To this end, we pair-wise compare subsets of joints (shown as the yellow dots in Fig. 2) for all interaction frames of two sequences. We iterate over all sequences and select the one with the minimal distance of the joint sets. We can efficiently identify the frame in each sequence with a 2D adaptation of Kabsch algorithm [26]. Based on the sequence with minimal distance, we label each other sequence as *prime* if the distance is below 0.5, and *inferior* otherwise.

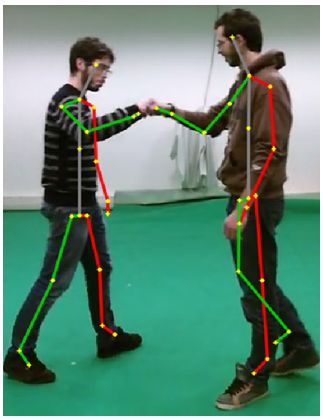


Fig. 2: Frame subwindow with superimposed pose data. Green: right side, red: left side.

Initial model learning We learn body part detectors w_i^j (Eq. 2) from the prime sequences. We determine, for each part, the type, spatial resolution and temporal extent. The spatial resolution indicates the cell size. For HOF, the temporal extent dictates how many frames around the epitome frame are used.

For each interaction, we train body part detectors for both persons using Dual Coordinate Descent SVM (DCD SVM) solvers [27]. After the positive optimization round, we perform a round of negative hard detection [18]. We can use different sources of negative examples. To train a model discriminatively we use a balanced selection of examples from all the other classes than the one being learned, as a source of negative examples. We can also train non-discriminatively by taking the negative examples from a completely different data

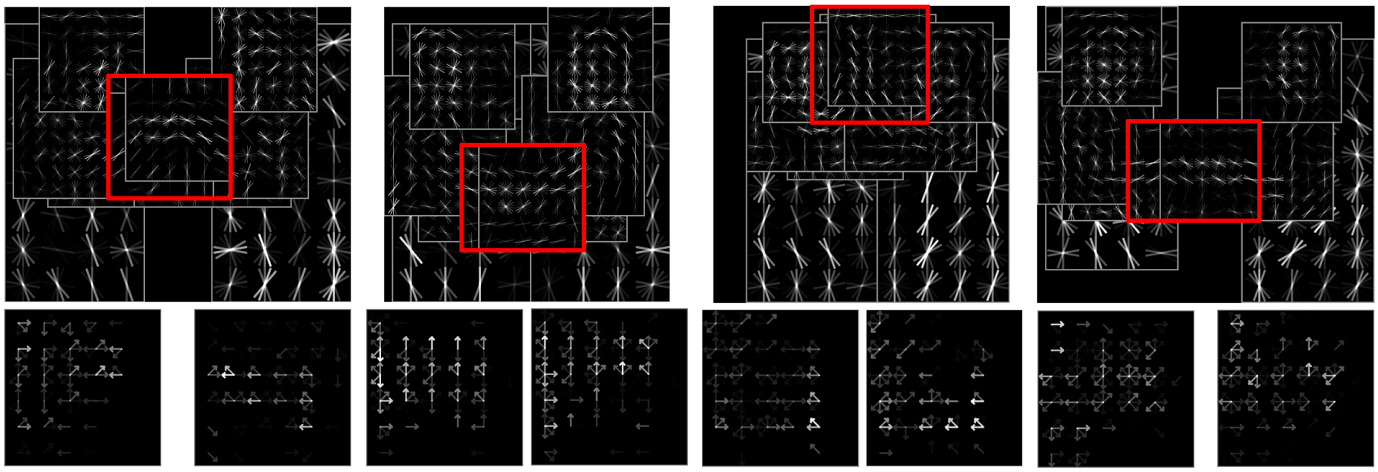


Fig. 3: Top row: HOG pose models for fist bump, hand shake, high five and pass object. Bottom row: HOF features of the right hands. The red rectangle indicates the enclosing bounding box of the two hands. Note the vertical hand movement for the hand shake model and the horizontal movement for fist bump.

source, to create realistic motion patches. This is beneficial to the model because it prevents overfitting to the environment of the training set. As a thrid option we can alternate between the discriminative and non-discriminative negative examples during training. This way we optimize both the subtle differences between features of different classes and we prevent overfitting to the environment.

Epitome and model refinement Once an initial model is constructed, we apply it to both prime and inferior training sequences of the particular class to detect new latent positive interaction examples. We search for the highest scoring frame in each sequence to add to the positive example set. Given that the initial epitome frames are selected solely based on pose, this step allows us to better represent the motion of the body. The resulting positive example set is used to optimize the model features and to determine all part biases and deformation parameters using the DCD SVM solvers. Example models are shown in Fig. 3.

C. Spatio-Temporal Localization

With a trained model, we can detect interactions in both space and time. We first detect interactions in individual frames, and then link these in time to form interaction tubes.

We generate a feature pyramid for each of the feature types to detect interactions at various scales. We extend the formulation to deal with feature types with a temporal extent, such as HOF. Based on Eq. 1, we generate a set of detection candidates spanning the entire video. In practice, we evaluate non-overlapping video segments. For a temporal HOF size of nine frames, we evaluate every ninth frame. Overlapping detections are removed with non-maximum suppression.

Interaction tubes We link frame detections into interaction tubes (see Fig. 4). To this end, we first sort candidate detections on detection score. Each tube starts with the best scoring detection. We consider the frame of this detection to be the epitome of the interaction. Then we greedily assign the detections of adjacent frames to the current tube. A detection is only added if it satisfies a minimum spatial overlap constraint ρ of 50% and a maximum area deviation of 50% with respect to the detection at the epitome. We iterate until all candidate detections have been assigned to a tube. Finally we remove all tubes with only a single detection.

IV. Experiments and Results

Given that we can use different sources of negative examples to train our models, we want to evaluate the accuracy of our model when localizing interactions that differ slightly. To address this scenario, we make use of a novel dataset: *ShakeFive2*. We train interaction detection models on this dataset and present the performance of different settings. Additionally, we test these models on the publicly available *UT-Interaction* [3] dataset.

ShakeFive2 consists of 94 videos with five close proximity interaction classes: *fist bump*, *hand shake*, *high five*, *hug* and *pass object*. Each video contains one two-person interaction, recorded under controlled settings but with small variations in

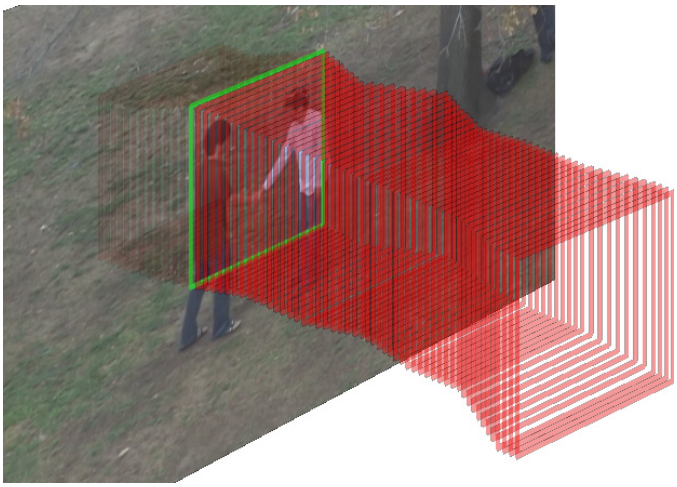


Fig. 4: Detected spatio-temporal interaction tube (red) for a hand shake. The green rectangle shows the best detection.

Fig. 5: Example frames from the datasets used in this paper: ShakeFive2 and UT-Interaction. Top row: hand shake, bottom row: hug.

viewpoint (Fig. 5). For each person in each frame, 2D joint position data obtained using Kinect2 is available.

UT-Interaction consists of two sets of 10 videos each. The first set features two persons in interaction per video, while the second set contains multiple pairs per video. The following interactions are performed: *hand shake*, *hug*, *kick*, *point*, *punch* and *push*. No pose data is available but bounding boxes are provided. To have a more tight estimate of the interaction per frame, we use the bounding box data from [10].

We have created three different experiments to test the effect of: (i) *non-discriminative training* (ND), where we harvest negative examples in random frames of the Hannah dataset [28]; (ii) *discriminative training* (D), where we use the examples of the other classes than the one we are training as the negative data source; (iii) *mixed training* (M), where we alternate between negative examples from Hannah and from training examples of the other classes. Data from Hannah helps extracting suitable motion patches because its visual environtment differs from the training data.

A. Performance Measurements

As we detect interactions in both space and time, we use the average intersection over union of the ground truth G and detected tube P as in [29]. G and P are two sets of bounding boxes and θ is the set of frames in which either P or G is not empty. The overlap is calculated as:

$$IoU(G,P) = \frac{1}{\|\theta\|} \sum_{f \in \theta} \frac{G_f \cap P_f}{G_f \cup P_f} \tag{3}$$

We evaluate different minimal overlap thresholds σ for which $IoU(G,P) \geq \sigma$. For cross-validation tests, we report the mean average precision (mAP) scores as the mean of the areas under the curves of each fold.

We consider two testing scenarios: single-class (SC) and multi-class (MC). For *single-class* detection, we apply a detector for a given interaction class to test videos of that class only. This scenario measures the spatio-temporal localization accuracy. In the *multi-class* scenario, we use the detector on all available test sequences in the dataset. This allows us to test for confusions with other interactions. In the multi-class scenario, the same interaction can be detected with models of different classes. This common situation will lead to false positives as we do not compare or filter these detections, but it gives a good indication of the mAP performance difference with and without discriminative training.

B. Results

For the three experiments we have conducted on the *ShakeFive2* dataset, we refer to the five interactions we evaluate as: FB (fist bump), HS (hand shake), HF (high five), HU (hug) and PO (pass object). In Table I we show the results. We note that on average the mixed training model performs best, though for some interaction classes, such as hand shake, the non-discriminative training model gives the best results.

When we compare the results of the non-discriminative training model to the discriminative and the mixed models, we can see in Fig. 6 that the mAP score decreases slowest with the mixed model, for increasing σ. This is the case in both the single class and multi-class scenarios.

As the mixed training model performs best in the first experiments we have tested the models that overlap with interactions from UT-Interaction on this data. The results of this experiment are shown in Table II. We can see that the performance differences between the hand shake model and the hug model are quite significant. We also note that Set #2 performs better than Set #1 one in our tests.

TABLE I: Non-discriminative (ND), Discriminative (D) and Mixed (M) mAP scores on *ShakeFive2* in a single-class (SC) and multi-class (MC) scenario.

ND/D/M	SC/MC	FB	HS	HF	HU	PO	**Avg.**
ND	SC	0.79	1.00	0.87	0.63	0.91	**0.84**
D	SC	0.97	0.85	0.80	0.47	0.75	**0.77**
M	SC	0.93	0.87	0.88	0.76	0.84	**0.86**
ND	MC	0.42	0.90	0.48	0.32	0.59	**0.54**
D	MC	0.71	0.80	0.76	0.45	0.53	**0.65**
M	MC	0.65	0.77	0.84	0.72	0.55	**0.71**

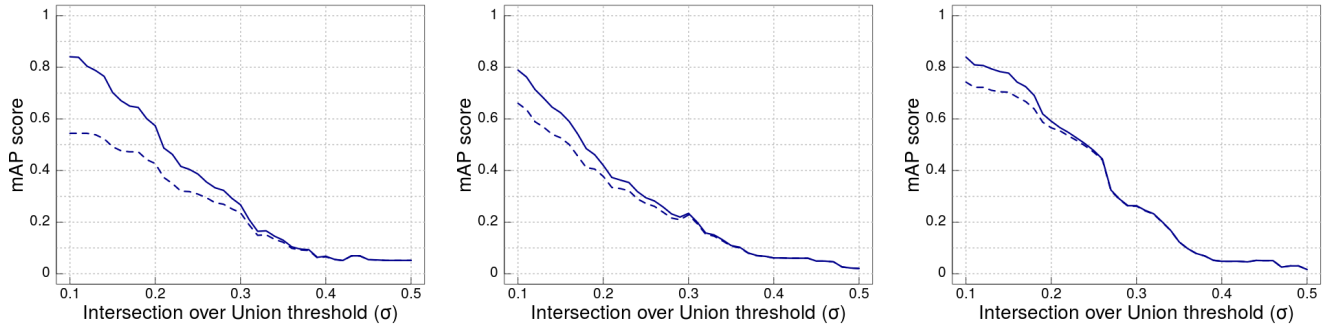
TABLE II: Single-class (SC) and multi-class (MC) mAP scores for UT-Interaction.

	Set	HS	HU	**Avg.**
SC	#1	0.60	0.21	**0.51**
	#2	0.82	0.44	
MC	#1	0.49	0.21	**0.48**
	#2	0.79	0.43	

V. Conclusion

We introduced a discriminatively trained interaction localization model. We have shown how to train it using multiple pose and motion features per part. The model's efficacy at localizing fine-grained interactions is shown on two challenging datasets containing human interactions with subtle differences.

We achieve a maximum mAP score of 0.86 for the mixed training model in the single class scenario for the experiments



(a) Non-discriminative training (b) Discriminative training (c) Mixed training

Fig. 6: 3-fold cross-validation mAP scores over all interaction classes in the single-class (solid line) and multi-class (dashed) scenarios of ShakeFive2 for increasing values of σ.

on ShakeFive2. The multi-class scenario achieves a maximum mAP score of 0.71 for the same model. On the UT-Interaction dataset we achieve an average performance of 0.51 and 0.48 in the single class and multi-class scenarios, respectively.

At this moment pose data is required to train our models. We would like drop this requirement by extending our model in future work. Another improvement would be modeling multiple perspectives to improve viewpoint independence.

REFERENCES

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proceedings International Conference on Pattern Recognition (ICPR)*, 2004, pp. 32–36.

[2] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," *Proceedings CVPR 2009*, 2009.

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

[4] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, 2011.

[5] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[6] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3192–3199.

[7] C. van Gemeren, R. T. Tan, R. Poppe, and R. C. Veltkamp, "Dyadic interaction detection from pose and flow," in *Proceedings Human Behavior Understanding Workshop (ECCV-HBU)*, 2014, pp. 101–115.

[8] A. Patron-Perez, M. Marszałek, I. Reid, and A. Zisserman, "Structured learning of human interactions in TV shows," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, no. 12, pp. 2441–2453, 2012.

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

[10] F. Sener and N. İkizler-Cinbis, "Two-person interaction recognition via spatial multiple instance embedding," *Journal of Visual Communication and Image Representation*, vol. 32, no. C, pp. 63–73, 2015.

[11] Y. S. Sefidgar, A. Vahdat, S. Se, and G. Mori, "Discriminative key-component models for interaction detection and recognition," *Computer Vision and Image Understanding (CVIU)*, vol. 135, pp. 16–30, 2015.

[12] I. Laptev, "On space-time interest points," *International Journal of Computer Vision (IJCV)*, vol. 64, no. 2-3, pp. 107–123, 2005.

[13] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[14] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision (IJCV)*, vol. 103, no. 1, pp. 60–79, 2013.

[15] B. Ni, P. Moulin, X. Yang, and S. Yan, "Motion part regularization: Improving action recognition via trajectory selection," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3698–3706.

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*. IEEE Computer Society, 2012, pp. 1290–1297.

[17] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 9, pp. 1775–1788, 2014.

[18] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[19] B. Yao, B. Nie, Z. Liu, and S.-C. Zhu, "Animated pose templates for modelling and detecting human actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 3, pp. 436–452, 2014.

[20] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2642–2649.

[21] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, 2008, pp. 995–1004.

[22] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 12, pp. 2878–2890, 2013.

[23] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Proceedings European Conference on Computer Vision (ECCV) - Part V*, 2010, pp. 168–181.

[24] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester, "Cascade object detection with deformable part models," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2241–2248.

[25] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 437–446.

[26] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 34, no. 5, pp. 827–828, 1978.

[27] J. S. Supancic III and D. Ramanan, "Self-paced learning for long-term tracking." in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2379–2386.

[28] A. Ozerov, J. Vigouroux, L. Chevallier, and P. Pérez, "On evaluating face tracks in movies," in *Proceedings International Conference on Image Processing (ICIP)*, 2013, pp. 3003–3007.

[29] J. C. van Gemert, M. Jain, E. Gati, and C. G. M. Snoek, "APT: Action localization proposals from dense trajectories," in *Proceedings British Machine Vision Conference (BMVC)*, 2015, p. A117.