Spatio-Temporal Detection of Fine-Grained Dyadic Human Interactions

Coert van Gemeren, Ronald Poppe, and Remco C. Veltkamp*

Interaction Technology Group, Department of Information and Computing Sciences,
Utrecht University, The Netherlands
{C.J.VanGemeren, R.W.Poppe, R.C.Veltkamp}@uu.nl

Abstract. We introduce a novel spatio-temporal deformable part model for offline detection of fine-grained interactions in video. One novelty of the model is that part detectors model the interacting individuals in a single graph that can contain different combinations of feature descriptors. This allows us to use both body pose and movement to model the coordination between two people in space and time. We evaluate the performance of our approach on novel and existing interaction datasets. When testing only on the target class, we achieve mean average precision scores of 0.82. When presented with distractor classes, the additional modelling of the motion of specific body parts significantly reduces the number of confusions. Cross-dataset tests demonstrate that our trained models generalize well to other settings.

Keywords: human behavior, interaction detection, spatio-temporal localization

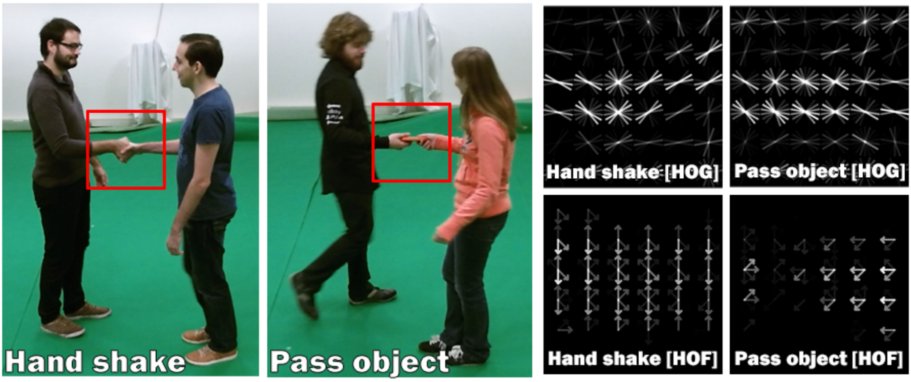


Fig. 1: Hand shake and object pass interactions with similar poses. We introduce a model to detect interactions that differ slightly in their spatio-temporal coordination by modeling pose and motion of specific body parts.

* This publication was supported by the Dutch national program COMMIT.

1 Introduction

Action recognition in videos continues to attract a significant amount of research attention [14]. Starting from the analysis of individuals performing particular actions in isolation (e.g. [19]), there is a trend towards the contextual analysis of people in action. There is much interest in the understanding of a person's actions and interactions in a social context, with research into the automated recognition of group actions [2] and human-human interactions [13,17].

This paper contributes to the latter category. We focus on two-person (*dyadic*) interactions such as shaking hands, passing objects or hugging. The type of interaction in which people engage informs us of their activity, the social and cultural setting and the relation between them. Automated detection of interactions can improve social surveillance, for example to differentiate between friendly and hostile interactions or to determine whether a person in an elderly home is a staff member, family member or unrelated visitor.

Poses of people in different interactions can be visually similar, for example when shaking hands or handing over an object (see Fig. 1). To differentiate between interactions, the *coordinated movement* of the people provides an additional cue. Not all body parts play an equally important role in each interaction. For example, a hand shake is characterized by the movement of the right hands. The distinction between such interactions requires a *fine-grained* analysis of the specific pose and body motion of both persons involved in the interaction.

In this paper, we detect dyadic interactions based on structural models [29] that combine pose (HOG) and movement (HOF) information. We train classifiers from videos and focus on those parts of the video that characterize the interaction. This enables us to distinguish between interactions that differ only slightly. An advantage of our method is that we can detect where the interaction occurs in a video in both space and time. This property allows us to identify who is involved in the interaction, or who hands over an object to whom.

Our *contributions* are as follows. First, we model the coordinated body movement of the people involved. We introduce a novel model to exploit these cues and to detect interactions in both space and time. Second, we present a procedure to train a detector from a few examples with pose information. Third, we demonstrate the performance of our framework on publicly available datasets. We report spatio-temporal localization performance for models trained only on the target interaction class.

We discuss related work in the next section. In Section 3, we introduce our model and detail the training and test procedures. The evaluation of our work appears in Section 4. We conclude in Section 5.

2 Related Work on Interaction Detection

The progress of vision-based action recognition algorithms is impressive [14]. Initial success was mainly based on bag-of-visual-word (BoVW) approaches that map image feature distributions to action labels [19]. Wang et al. [26] link these features over time into dense trajectories, allowing for more robust representations of movement. The work has been extended by clustering the trajectories to enable the *spatio-temporal* detection of actions [25].

While these representations have achieved state-of-the-art performance, they do not explicitly link image features to human body parts. The availability of body pose and, especially, body movement information has been found to increase action classification performance [5]. This is because the pose or movement of some body parts is often characteristic. For example, arm movement is more discriminative than leg movement in a hand shake. Without pose information, discriminative patterns of image movement can only be modeled implicitly, e.g. using clusters of dense trajectories [11] or co-occurring spatio-temporal words [32]. These approaches are automatic but less reliable in the presence of other motions, when multiple people interact with each other in close proximity.

Part-based models such as Deformable Part Models (DPM, [3]) and poselets [1] can detect people in an image and localize their body parts. These models employ body part detectors and impose spatial contraints between these parts. DPMs are sufficiently flexible to describe articulations of the body [29]. This enables the detection of key poses representative of an action [15]. Often, two actions cannot be distinguished based on a single key pose, see Fig. 1. Movement can then be used to distinguish between classes [23]. Yao et al. [30] represent actions as a combination of a pose and a mixture of motion templates.

In this paper, we follow this line of research, but extend it to the detection of interactions. Researchers have started to analyze behavior of multiple people [2,9]. Here, we focus on the recognition of two-person interactions. Recent work in this area has used gross body movement and proximity cues for the detection of interactions. A common approach is to first detect faces or bodies using off-the-shelf detectors [13,18]. Detections of individuals can be paired and the resulting bounding volume can be used to pool features in a BoVW approach [10].

The relative distances and orientations between people can also be used to characterize interactions. Patron-Perez et al. [13] use coarse distance labels (e.g., far, overlap) and differences in head orientation. They also include local features around each person such as histograms of oriented gradients (HOG) and flow (HOF). Sener and İkizler [21] take a similar approach but cast the training as multiple-instance learning, as not all frames in an interaction are considered informative. For the same reason, Sefidgar et al. [20] extract discriminative key frames and consider their relative distance and timing within the interaction.

Kong and Fu [7] observe that not all body parts contribute equally. Their method pools BoVW responses in a coarse grid. This allows them to identify specific motion patterns relative to a person's location but the level of detail of the analysis is limited by the granularity of the patches and the accuracy of the person detector. Yang et al. [28] found that a sequential approach of first detecting individuals and then recognizing their interaction does not perform well when there is physical contact. They significantly improve classification performance by building detectors for various types of physical interactions such as hand-hand

and hand-shoulder touches. Here we also focus on physical interactions, but we look at the *fine-grained* differences between visually similar classes.

Proximity and orientation are good cues for detection of coarse interaction classes, but less so to detect fine-grained interactions such as those in social encounters. These are characterized by body movements that are visually similar, but differ slightly in the temporal coordination. To distinguish between such interactions, we need to more effectively model the coordination between the people involved.

Kong et al. [8] train detectors for attributes such as “outstretched hands” and “leaning forward torso” and consider their co-occurrences. Given sufficiently detailed attributes, fine-grained interactions could be detected. However, as each detector is applied independently, false detections are likely to occur. Van Gemeren et al. [24] use interaction-specific DPMs to locate people in characteristic poses. They then describe the coordinated movement in the region in between DPM detections. 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 this issue by combining the detection of the people and their interaction in a single step. We diverge from Yao et al. [30], by constraining how pose and motion are coordinated in a dyadic scenario, so we can model spatio-temporal coordination at a much more fine-grained level. Yao et al. train and test their model on human-object interaction tasks, whereas we focus specifically on dyadic human interactions.

3 Modeling Fine-Grained Coordinated Interactions

We model two-person interactions based on DPMs for pose recognition in images, introduced by Yang and Ramanan [29]. We solve three limitations. First, parts are not locally centered on body joints but are specific for an interaction and typically encode the relative position and articulation of a body part, similar to poselets [1]. Second, we allow each part detector to contain multiple image cues. We explicitly enable the combination of static and temporal features. We can thus decide per body part whether pose, motion or a combination is most discriminative for a specific interaction. Third, we consider two persons simultaneously. Our formulation models the spatial and temporal coordination between their poses and movements at a fine scale. We discuss the model, training algorithm and detection procedure subsequently.

3.1 Model Formulation

Our model is motivated by the observation that many interactions are characterized by a moment where the poses of two people are spatially coordinated and the movement of a specific part of the body is temporally coordinated.

Let us define graph $G = (V, E)$, with V a set of K body parts and E the set of connections between pairs of parts [29]. Each body part i is centered on location $l_i = (x_i, y_i)$. For clarity, we omit in our formulation the extent of the body part's area, as well as scaling due to processing an image i at multiple resolutions. The scoring for a part configuration in image I is given by:

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

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 [29]. These distances are defined with respect to root factor r, which allows for scaling of parts with a different cell resolution as the root part [3]. w_{ij} encodes the rest location and the rigidity of the connections between parts.

We now describe our adaptations of this model for the modeling of fine-grained dyadic interactions.

Class-specific part detectors While [29] considers different body part orientations as parameters in the model, we learn class-specific detectors that encode the articulation of the body part directly. Though our method allows for modeling multiple mixtures per part, our data only features homogeneous interactions recorded from a specific viewing angle. Therefore, we use only a single detector per class, instead of a mixture of part detectors. Aside from having data that features interactions performed in different ways from multiple viewpoints, increasing the amount of mixtures would also require a larger amount of samples.

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:

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

$\phi_i^j(I, l_i)$ denotes a feature vector of type j (e.g., HOG or HOF) 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. As such, our formulation is different from Yao et al. [30], 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. To our knowledge currently no methods exist that model dyadic interactions as a single part based model.

In the experiments presented in this paper, the sub-tree of each person has a *torso* root node with four child parts: *head*, *right upper arm*, *right lower arm* and *right hand*.

3.2 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 dyadic interaction. The metadata of the training videos contains 3D skeleton joint positions, from which we calculate 2D projections. We use this to place parts on limb locations. We assume the sequences are segmented in time to contain the interaction of interest. As the temporal segmentation relies on human annotations the start and the end of an interaction are not precisely marked. Therefore we consider a single short sequence of frames most representative for the interaction in each sequence, as the base of the model. We call this sequence the *epitome*. We guarantee that the epitome is taken from the temporally segmented sequence.

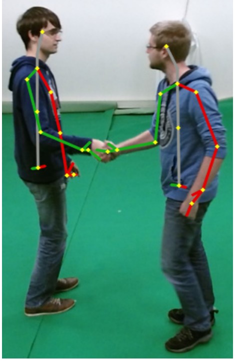


Fig. 2: Frame with superimposed pose data.

Training consists of 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. To this end, we pair-wise compare the joint sets of all frames in two sequences. For our experiments, we consider all joints in the right arm of both persons in interaction (green parts in Fig. 2). We can efficiently identify the epitome in each sequence with the Kabsch algorithm [6]. We use it to compare sets of coordinates in a translation, scale and rotationally invariant way. Based on the Kabsch distance between the video with the lowest sum distance to all other videos, we label each sequence as *prime* if this distance is below 0.5, and *inferior* otherwise. Essentially we separate the videos in which the skeleton poses look-alike, from the videos where they don't.

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. In this paper, we consider HOG and HOF features [26] but the DPM inference algorithm is well suited to incorporate a learned feature extractor such as convolutional neural networks (CNN) [4]. 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 [22]. After the positive optimization round, we perform a round of hard negative detection [3]. Negative examples are harvested in random frames of the Hannah dataset [12], to avoid overfitting to a particular training set, and to allow for the extraction of realistic motion patches. After optimizing all part mixtures, we combine all parts into a single spatio-temporal DPM (SDPM). The locations of the parts are based on the average relative center locations in the pose data.

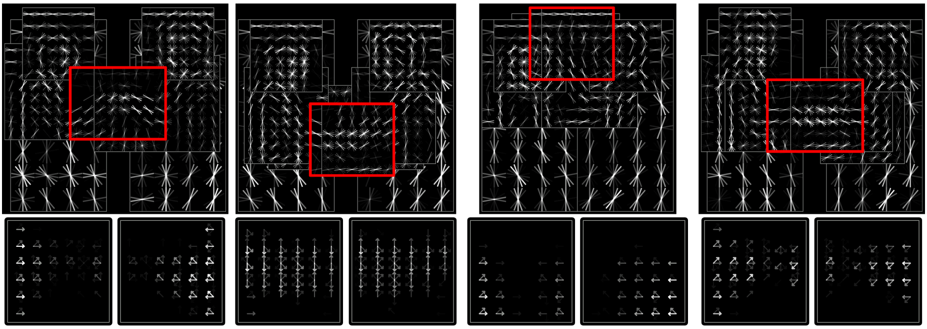


Fig. 3: Top row: HOG pose models for fist bump, hand shake, high five and pass object. Bottom row: HOF features of the right hands. The red rectangle indicates the enclosing bounding box of the two hands.

Epitome and model refinement Once an initial SDPM is constructed, we apply it to both prime and inferior training sequences to detect new latent positive interaction examples. We search for the highest scoring frame in each sequence to update 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. Note the vertical hand movement for the hand shake model and the horizontal movement for fist bump.

3.3 Spatio-Temporal Localization

With a trained SDPM, we can detect interactions in both space and time. We specifically avoid 3D feature extraction during training because we want to be able to apply our model on data that does not contain any depth information. We first detect interactions in frame sequences that last shorter than a second, and then link these to form interaction tubes, without the use of depth information.

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. Based on Eq. 1, we generate a set of detection candidates spanning the entire video. In practice, we evaluate non-overlapping

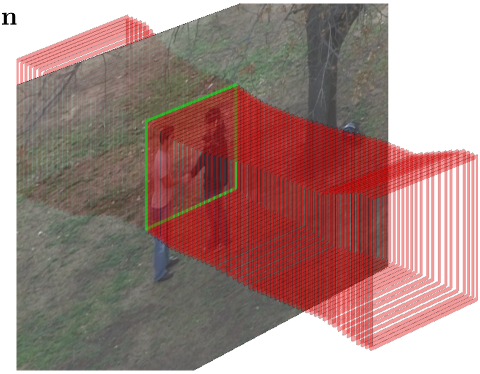


Fig. 4: Detected spatio-temporal interaction tube (red) for a hand shake. The green rectangle shows the best detection.

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). We sort candidate detections on detection score. Each tube starts with the best scoring detection. We then 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 best detection. We iterate until all candidate detections have been assigned to a tube. Finally we remove all tubes with only a single detection.

4 Experiments and Results

Previous research on interaction *recognition* has considered assigning labels to video sequences that have been segmented in both space and time. In contrast, we focus on *spatio-temporal detection* of interactions from unsegmented videos. To address this scenario, we present a novel dataset and our performance measures. Subsequently, we summarize the setup and results of our experiments.

4.1 Datasets

As available interaction datasets contain behaviors that are visually quite dissimilar, we introduce a novel dataset *ShakeFive2*[1] with interactions that differ slightly in their coordination. We train interaction detection models on this dataset and present the performance of different settings. In addition, we test these models on publicly available interaction datasets *SBU Kinect* [31] and *UT-Interaction* [17]. Example frames from each of these datasets can be seen in Fig. 5.

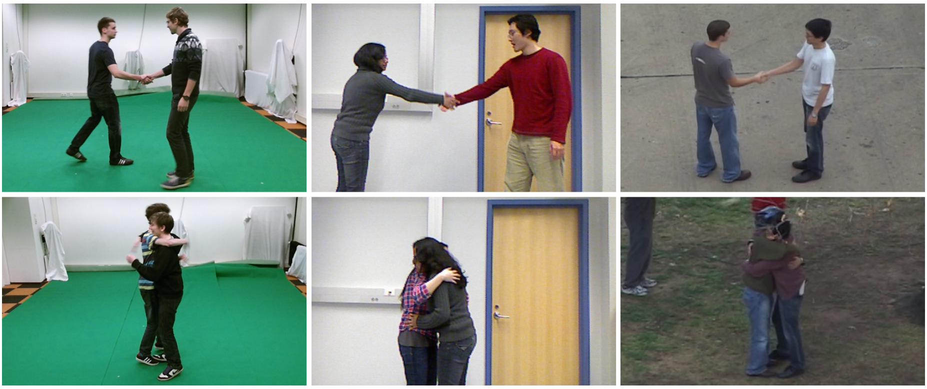


Fig. 5: Example frames from the datasets used in this paper: ShakeFive2, SBU Kinect and UT-Interaction. Top row: hand shake, bottom row: hug.

[1] *ShakeFive2* is publicly available from `https://goo.gl/ObHv36`

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 viewpoint. We note that in the pass object interaction a small orange object is passed from one person to the other. This is the same small object for all videos. For each person in each frame, 3D joint position data obtained using Kinect2 is available.

SBU Kinect involves two actors performing one interaction per video in an indoors setting. The interactions are: *hand shake*, *high five*, *hug*, *pass object*, *kick*, *leave*, *punch* and *push*. Pose data, obtained with a Kinect, is provided but not always accurate. From the 260 videos, we exclude 42 with incorrect pose data.

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. These span the entire spatial extent of the interaction. To have a more tight estimate of the interaction per frame, we use the bounding box data from [21].

4.2 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 [25]. 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 create one precision-recall diagram per fold. 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 test the detector on all available test sequences in the dataset. This allows us to determine whether there are 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. The reported mAP scores are therefore conservative but demonstrate the performance of our models without discriminative training.

To assess which pairs of classes are more often confused, we introduce a novel measure that takes into account the spatio-temporal nature of our problem. We test a trained detector in the single-class and multi-class detection scenarios and calculate the difference in mAP (d-mAP) scores between these two settings.

When no false positives have been identified, the d-mAP score is zero. Higher d-mAP scores are due to the performance loss caused by the false positives for the particular distractor class.

4.3 Features and Experiment Setup

Our model can be trained using different types of descriptors per part. In our experiments, we consider HOG and HOF descriptors. For HOG, we use the gradient description method of [3], which differs slightly from [26]. Optical flow is calculated with DeepFlow [27]. For the time dimension of HOF, we use three bins of three frames each. For a 30fps video, this covers about a third of a second.

We use a **HOG** model that describes the torso with 4×8 cells, the right upper arm with 7×8, right lower arm with 9×7 and the right hand and head with 6×6 cells. The number of pixels per cell is 8×8 for the torso and 4×4 for other body parts. The **HOF** model is similar but all body parts are encoded as HOF. The **HOGHOF** model describes the torso and head as HOG, the right upper and lower arms as HOG and HOF and the right hand with HOF.

Models are trained on the data of ShakeFive2 using three-fold cross-validation. In each fold, there are six or seven sequences per class. We therefore train on either 12 or 13 sequences only. The performance in the single-class scenario is calculated as the average performance over the three folds. In the multi-class scenario, we combine the test folds of the different interaction classes, creating a set of 30–34 videos of which six or seven are of the target class.

4.4 Detection Results

We first investigate the added value of using motion information for interaction detection. We test the **HOG**, **HOF** and **HOGHOF** models on the ShakeFive2 dataset. We refer to the five interactions as FB (fist bump), HS (hand shake), HF (high five), HU (hug) and PO (pass object). Results for the single-class (SC) and multi-class (MC) scenarios are shown in Table 1. We use a minimal overlap σ between the detected tube and the ground truth volume (Eq. 3) of 10%.

Table 1: Single-class (SC) and multi-class (MC) mAP scores on ShakeFive2.

	SC/MC	FB	HS	HF	HU	PO	**Avg.**
HOG	SC	0.74	0.79	0.75	0.61	0.95	**0.77**
HOF	SC	0.55	0.75	0.70	0.65	0.55	**0.64**
HOGHOF	SC	0.83	0.95	0.83	0.61	0.88	**0.82**
HOG	MC	0.32	0.55	0.39	0.37	0.63	**0.45**
HOF	MC	0.23	0.60	0.48	0.51	0.28	**0.42**
HOGHOF	MC	0.54	0.88	0.50	0.34	0.57	**0.57**

When tested on only videos of the same class (SC), we see that the **HOGHOF** model outperforms both **HOG** and **HOF**. This demonstrates that interactions are most accurately detected by a combination of pose and motion information. The lower performance of **HOF** indicates that movement information alone is not sufficient to robustly detect interactions from video. When additional sequences of other interaction classes are available (MC), we notice a significant drop for all models but less so for **HOGHOF**. Especially the lack of pose information in the **HOF** model appears to cause misclassifications between interactions. The combination of pose and motion in the **HOGHOF** model appears to work best. Note that all models are trained on at most 13 positive training sequences and that the other interactions are not provided as negative samples. The models are therefore not trained to discriminate between interaction classes.

Table 2: d-mAP scores for the **HOG** (left) and **HOGHOF** (right) models on ShakeFive2. In columns the true class, in rows the tested class.

	FB	HS	HF	HU	PO
FB	▨	0.41	0.24	0.16	0.44
HS	0.22	▨	0.15	0.15	0.31
HF	0.32	0.31	▨	0.20	0.25
HU	0.23	0.26	0.23	▨	0.19
PO	0.15	0.27	0.07	0.05	▨

	FB	HS	HF	HU	PO
FB	▨	0.19	0.16	0.13	0.28
HS	0.04	▨	0.04	0.04	0.09
HF	0.26	0.19	▨	0.11	0.16
HU	0.29	0.18	0.24	▨	0.22
PO	0.19	0.25	0.09	0.05	▨

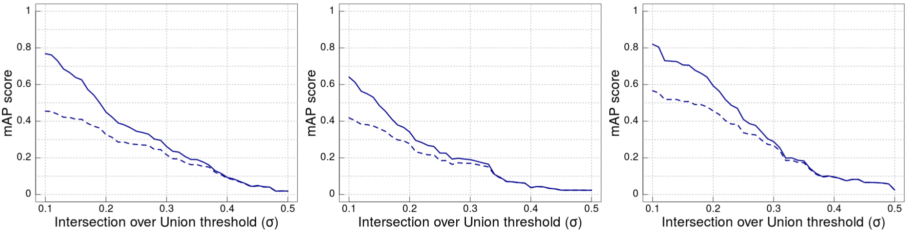
There are some differences in performance between the interaction classes. Hand shakes can be detected relatively robustly by all models, whereas especially hugs are often not detected. In the multi-class setting, we can investigate how often interaction classes are confused. We present the d-mAP multi-class detection scores on ShakeFive2 for the **HOG** and **HOGHOF** models in Table 2. For the **HOG** model, there are many confusions. Apparently, the pose information alone is not sufficiently informative to distinguish between interactions that differ slightly in temporal coordination: hand shake, fist bump and pass object. The number of confusions for the **HOGHOF** model is much lower. The additional motion information can be used to reduce the number of misclassification between visually similar interactions.

We note that especially fist bump and hand shake have fewer confusions with the **HOGHOF** model compared to the **HOG** model. However, the **HOGHOF** model for pass object has more confusions. We expect that the variation in the performance of this interaction leads to a suboptimal model during training. This can be seen in Fig. 3 as well. The HOG description of the pose is somewhat ambiguous, while the HOF descriptor of the hands is similar for the pass object and fist bump interactions. Indeed, many pass object interactions are detected as fist bumps.

4.5 Parameter Settings

Next, we investigate the influence on the detection performance of the most important parameters of our models: the minimal tube overlap (σ), the minimal spatial overlap (ρ) and the number of training sequences.

Minimal tube overlap is a measure of how accurate the detections are in both space and time. A higher threshold σ requires more accurate detection. In line with [25], we vary this threshold from 0.1 to 0.5. Fig. 6 shows the performance of the three models for increasing σ. We note that **HOG** (Fig. 6a) shows a better performance than **HOF** (Fig. 6b) when σ increases. When HOG and HOF are combined (**HOGHOF**, in Fig. 6c), we observe a significant increase in performance and mAP scores that remain higher for larger values of σ.



(a) **HOG** (b) **HOF** (c) **HOGHOF**

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

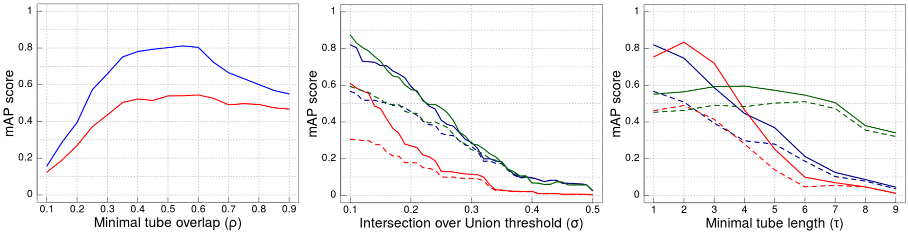
Minimal spatial overlap Subsequent detections in time are linked provided that they sufficiently overlap spatially. The default threshold ρ of 50% is in line with object detection research but Fig. 7a shows the mAP scores for different values of ρ, with best results for $\rho = 58\%$. A higher value for ρ results in fewer links and, consequently, smaller tubes. With a lower threshold, noisy detections are more often linked to the tube, also resulting in a lower mAP.

Amount of training data We noticed that the **HOGHOF** models achieve good detection performance despite being trained on a small number of example sequences. Here we test the performance of the model when trained on different numbers of sequences. Fig. 7b shows the mAP scores when training on 2, 12–13 (3 folds), and 15–16 (6 folds) sequences. For the first setting, we sampled pairs of training sequences. Clearly, performance is lower when training on just two training sequences. The difference between 12–13 and 15–16 sequences is very small. This suggests that saturation occurs at a very low number of training data. This is advantageous as obtaining training sequences with associated pose data might be difficult, especially when many interaction classes are considered.

4.6 Performance on SBU Kinect and UT-Interaction

To compare our method to previous work, we also evaluate the performance on publicly available interaction datasets SBU Kinect and UT-Interaction. We train **HOGHOF** models on all available sequences in ShakeFive2. Results reported are for *cross-dataset* evaluation. In the single-class scenario, we only report the interactions are shared between ShakeFive2 and the other two datasets. We evaluate all available videos in the dataset in the multi-class scenario.

Even though the three datasets are similar in the type of interaction, there are several notable differences. First, there is variation between the datasets



(a) mAP per ρ (b) mAP and # videos (c) mAP per τ

Fig. 7: mAP scores for different parameter settings in the single-class (solid line) and multi-class (dashed) scenarios. Fig. 7a shows the influence of the minimal spatial overlap on the performance. Fig. 7b shows the performance with different amounts of training videos: 2 (red), 12-13 (blue) or 15-16 (green). Fig. 7c shows the influence on the minimal tube overlap for different datasets: ShakeFive2 (blue), SBU Kinect (red) and UT-Interaction (green).

in the viewpoint and the performance of the interactions (see also Fig. 5). For example, the average durations of hand shakes in ShakeFive2 and UT-Interaction are 27 and 100 frames, respectively, both at 30 frames per second. Also, the percentage of positive interaction frames differs. For UT-Interaction, 5% of the frames contain the interaction of interest. This is 12% for ShakeFive2, and all frames of SBU Interact contain the target interaction.

To account for differences in interaction length, we introduce minimal tube length τ. Tubes shorter than τ segments are removed. This is beneficial for datasets with significantly longer interactions than in the training data. Fig. 7c summarizes the performance of the **HOGHOF** model on the evaluated datasets. ShakeFive2 and SBU Kinect have similar profiles, UT-Interaction scores better for τ values around 4. For SBU Kinect and UT-Interaction, we set $\tau = 2$.

Table 3: Single-class (SC) and multi-class (MC) mAP scores for SBU Kinect.

	SC/MC	HS	HU	PO
HOGHOF	SC	0.94	0.68	0.87
HOGHOF	MC	0.71	0.53	0.24

SBU Kinect Table 3 summarizes the performance on SBU Kinect. We have tested the "noisy" variation of this dataset using our **HOGHOF** model with $\sigma = 0.1$, $\rho = 0.5$ and $\tau = 2$. We observe high scores in the single-class scenario, even though we did not train on this dataset. For comparison, Yun et al. [31] report classification performance on the dataset when using the pose features. They obtain 75%, 61% and 85% recognition accuracy for the hand shake, hug and pass object interactions, respectively. While these scores cannot be compared directly, it is clear that classification of segmented sequences already presents challenges. Detecting the interaction in space and time adds to the challenge.

Table 4: d-mAP scores for the **HOGHOF** models on SBU Kinect. In columns the true class, in rows the tested class.

	HS	HU	KI	LV	PC	PS	PO
HS	▨	0.03	0.08	0.06	0.12	0.18	0.14
HU	0.18	▨	0.21	0.14	0.22	0.24	0.26
PO	0.38	0.08	0.22	0.29	0.23	0.40	▨

We note that the detection of the pass object interaction scores particularly low in the multi-class setting compared to the single-class setting. To analyze confusions, Table 4 presents d-mAP values for all SBU Kinect interactions: hand shake (HS), hug (HU), kick (KI), leave (LV), punch (PC), push (PS) and pass object (PO). Many hand shake and push interactions are detected as pass object. These three interactions are characterized by extended, horizontally moving arms. The pass object model clearly is not discriminative enough to pick up on the subtle differences between the interactions.

Table 5: Single-class (SC) and multi-class (MC) mAP scores for UT-Interaction (left). Classification accuracies reported on UT-Interaction (right).

	Set	HS	HU	**Avg.**
SC	#1	0.61	0.39	**0.57**
	#2	0.90	0.36	
MC	#1	0.48	0.38	**0.46**
	#2	0.63	0.36	

Method	**Avg.**
Raptis & Sigal [15]	100%
Ryoo [16]	85%
Sener & İkizler [21]	100%
Zhang, et al. [32]	100%

UT-Interaction Finally, we evaluate the **HOGHOF** models on the UT-Interaction dataset. Results of our model and previously reported results are summarized in Table 5. A direct comparison with other works is difficult for a number of reasons. First, we report detection results only for hand shake and hug, the common interactions between ShakeFive2 and UT-Interaction. Second, we report spatio-temporal localization results, whereas other works consider a recognition scenario. In this setting, volumes segmented in space and time are classified. Third, we train our models on a different dataset.

Table 5 shows the detection results on both sets of UT-Interaction. Our **HOGHOF** can detect multiple simultaneous interaction, as witnessed by the scores on set 2. The detection of hugs is much lower. We attribute this to the longer duration of the hugs. Many hugs are not detected for a sufficient number of subsequent frames. As a result, there are missed detections. Higher values for τ can alleviate this problem.

5 Conclusions and Future Work

We have introduced a novel model for the detection of two-person interactions. Our spatio-temporal deformable part models combine pose and motion in such a way that we can model the fine-grained coordination of specific body parts. For the first time, we address the spatio-temporal detection of interactions from unsegmented video. Our approach allows us not only to say whether an interaction has occurred, but also to recover its spatial and temporal extent.

Interaction models are trained from only a few videos with pose information. On the novel ShakeFive2 dataset, we achieve mAP scores of 0.82 when training on 12–13 sequences. In the presence of visually similar interactions, motion information reduces the number of misclassifications. We obtain mAP scores of 0.57 without discriminative training, and without filtering the detections. Moreover, our cross-dataset evaluations on the publicly available UT-Interaction and SBU Kinect datasets demonstrate that the model generalizes to different settings.

Despite its good performance, the method has some limitations. Most importantly, the number of false detections is considerable. Currently, we can have several detections of the same interaction. By filtering these, we can reduce the number of false positives. This will allow us to report classification results. Another improvement is the discriminative training of the interaction models. This is likely to improve the detection performance as each model can focus on those parts of the pose or motion that are discriminative for the interaction.

Pose data is required to train our models. We are considering incremental training schemes that alleviate this need. Finally, we would like to include multiple perspectives to improve viewpoint independence. While there is some variation within and between the datasets that we have evaluated, viewpoint invariance will further increase the applicability of our work.

Together, we envision that these improvements can bring closer the automated spatio-temporal detection of a broad range of social interactions in unconstrained video material. This will allow for the automated analysis of TV footage and web videos. Moreover, we aim at the application of our work in dedicated social surveillance settings such as in public meeting places and elderly homes.

References

1. 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*, pages 168–181, 2010.
2. W. Choi and S. Savarese. Understanding collective activities of people from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(6):1242–1257, 2014.
3. 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)*, 32(9):1627–1645, 2010.
4. 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)*, pages 437–446, 2015.
5. 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)*, pages 3192–3199, 2013.

6. W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978.
7. Y. Kong and Y. Fu. Close human interaction recognition using patch-aware models. *IEEE Transactions on Image Processing (TIP)*, 25(1):167–178, 2015.
8. Y. Kong, Y. Jia, and Y. Fu. Interactive phrases: Semantic descriptions for human interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(9):1775–1788, 2014.
9. T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(8):1549–1562, 2012.
10. M. J. Marín-Jiménez, E. Yeguas, and N. Pérez de la Blanca. Exploring STIP-based models for recognizing human interactions in TV videos. *Pattern Recognition Letters*, 34(15):1819–1828, 2013.
11. 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)*, pages 3698–3706, 2015.
12. A. Ozerov, J. Vigouroux, L. Chevallier, and P. Pérez. On evaluating face tracks in movies. In *Proceedings International Conference on Image Processing (ICIP)*, pages 3003–3007, 2013.
13. 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)*, 34(12):2441–2453, 2012.
14. R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.
15. M. Raptis and L. Sigal. Poselet key-framing: A model for human activity recognition. In *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2650–2657, 2013.
16. M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, pages 1036–1043, 2011.
17. 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.
18. M. S. Ryoo and J. K. Aggarwal. Stochastic representation and recognition of high-level group activities. *International Journal of Computer Vision (IJCV)*, 93(2):183–200, 2011.
19. C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proceedings International Conference on Pattern Recognition (ICPR)*, pages 32–36, 2004.
20. 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)*, 135:16–30, 2015.
21. F. Sener and N. İkizler-Cinbis. Two-person interaction recognition via spatial multiple instance embedding. *Journal of Visual Communication and Image Representation*, 32(C):63–73, 2015.
22. J. S. Supancic III and D. Ramanan. Self-paced learning for long-term tracking. In *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2379–2386, 2013.
23. Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2649, 2013.

24. 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)*, pages 101–115, 2014.
25. 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)*, page A117, 2015.
26. 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)*, 103(1):60–79, 2013.
27. P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, pages 1385–1392, 2013.
28. Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3522–3529, 2012.
29. Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(12):2878–2890, 2013.
30. 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)*, 36(3):436–452, 2014.
31. 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 *Proceedings Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 28–35, 2012.
32. Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen. Spatio-temporal phrases for activity recognition. In *Proceedings European Conference on Computer Vision (ECCV)*, pages 707–721, 2012.