Dyadic Interaction Detection from Pose and Flow

Coert van Gemeren[1], Robby T. Tan[2],
Ronald Poppe[1], and Remco C. Veltkamp[1,⋆]

[1] Interaction Technology Group, Department of Information and Computing Sciences, Utrecht University, The Netherlands
[2] School of Science and Technology, SIM University, Singapore
{C.J.VanGemeren,R.W.Poppe,R.C.Veltkamp}@uu.nl, RobbyTan@unisim.edu.sg

Abstract. We propose a method for detecting dyadic interactions: fine-grained, coordinated interactions between two people. Our model is capable of recognizing interactions such as a hand shake or a high five, and locating them in time and space. At the core of our method is a pictorial structures model that additionally takes into account the fine-grained movements around the joints of interest during the interaction. Compared to a bag-of-words approach, our method not only allows us to detect the specific type of actions more accurately, but it also provides the specific location of the interaction. The model is trained with both video data and body joint estimates obtained from Kinect. During testing, only video data is required. To demonstrate the efficacy of our approach, we introduce the *ShakeFive* dataset that consists of videos and Kinect data of hand shake and high five interactions. On this dataset, we obtain a mean average precision of 49.56%, outperforming a bag-of-words approach by 23.32%. We further demonstrate that the model can be learned from just a few interactions.

1 Introduction

In the past years, a lot of progress has been made in recognizing human actions from video [1]. Initial research has mainly taken a holistic approach, modeling the area of interest in the video as a feature. Bag-of-word approaches have become popular and suitable for the distinction of broad categories of actions such as running and jumping [2, 3]. By generalizing over specific poses, viewpoints and person appearances, they have been found to be fitting to data "in the wild". However, this generalization eventually hampers the use of these holistic approaches for the detection of more fine-grained actions.

Another approach to action recognition is to first estimate the configuration of the body, and then use this representation for subsequent action recognition. While classification can typically be performed more effectively, the challenge of dealing with less-controlled videos is moved to the body pose estimation process.

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

H.S. Park et al. (Eds.): HBU 2014, LNCS 8749, pp. 101–115, 2014.

Recently, a lot of progress has been made. Notably, the introduction of pictorial structure models [4–6] and poselets [7] have paved the way for robust estimation of body poses in "in the wild". One challenge that remains is how to deal with the variations in the performance of an action. The same action can be performed in many ways and still be perceived as the same action. For example, sitting down will be performed differently depending on whether the person will sit on a barstool or on a chair. Another issue is that not all body parts contribute equally in the performance of an arbitrary action. For some actions (e.g. pointing), the position of the legs is not that important. This is also typically the case when a person interacts with another person or with the environment.

In this paper, we focus on the detection and recognition of dyadic interactions from video. These interactions are coordinated in a sense that the movement of one person depends on the movement of the other, and *vice versa*. For example, a hand shake requires both people to face each other, extend their arms forward, grab the other's hand and simultaneously move the hands up and down. If the shake of the hands would not be coordinated, we would probably not perceive the movements as a hand shake. Recognizing coordinated interactions therefore does not only require the detection of the movement of both persons individually, but also the assessment of the coordination of their movements. This notion is reflected in the approach proposed in this paper.

We start by detecting individual actions using poselets, which are templates that encode the specific configuration of parts of the body. Poselets are typically described using histograms of oriented gradients (HOG). Maji *et al.* [8] have used poselets for the recognition of actions from a single frame. In contrast, we consider videos and take advantage of the movement information by including histograms of oriented flow (HOF) information into the poselet representation, as in [9]. Yao *et al.* [10] also use a combination of HOG and HOF to recognize actions. They use a grammar-like representation in which the HOF determines transitions between different poses, encoded using HOG. In this formulation, temporal variations in the performance of an action can be overcome.

In this work, we estimate the locations of key joints from the poselet detections. The relative positions are learned during training using 2D joint positions estimated by Kinect. When the estimated joint positions of two persons are close, we further investigate the area. For example, to detect a hand shake, the right hand joints are the key joints and the overlapping area contains both hands. We encode this area with a combination of HOG and HOF and train a classifier for the interaction. This allows us to analyze the interaction at a more fine-grained level. During testing, only video data is used, making the approach suitable for the detection of dyadic actions in a wide range of applications, from surveillance to the fine-grained analysis of people in conversation. We evaluate our approach on *ShakeFive*, a novel dataset containing hand shakes and high fives, as well as the metadata gathered from Kinect.

Our model assumes that the scene is recorded by a static camera and actions are viewed from the side. We also assume that no occlusions occur in recording the interaction. We restrict ourselves in this manner because we are first and

foremost interested in the the ability to learn pictorial structures using Kinect, while recognizing the interactions without the use of hardware other than a standard camera.

We make the following contributions: First, we train a dyadic interaction recognition model using Kinect information in a controlled environment during training, while using the model to recognize dyadic interactions from the video data only. Second, we demonstrate that fine-grained analysis of the interaction around key joints is beneficial to the classification of these interactions. Finally, we provide a new dataset containing video and joint poses from Kinect of two individuals involved in a hand shake or a high five.

The paper is structured as follows. First, we discuss related work. In Section 3, we present our approach. We outline and discuss our experiments in Section 4 and conclude in Section 5.

2 Related Work

One popular approach to recognize actions from video is based on a bag-of-words representation. Distinct features based on edges, motion or both are found and collected within spatio-temporal regions of a video. While the recent trend to rely on trajectories of these points has shown potential, there is typically no connection between the features and the articulated human body. As a consequence, it is difficult to distinguish between actions that differ slightly. We can imagine, for instance, that running a hand through the hair is indistinguishable from scratching the head even though both are performed in different contexts. We therefore focus here on approaches that consider body poses as an intermediary level, to allow a more fine-grained analysis. We discuss recent work in this area, with a specific focus on their application for the classification of interactions between people.

While body poses are a convenient representation to learn actions from, there is typically a challenge in recovering these body poses from video. Body part models that encode both the appearance of individual body parts and the spatial relations between them have shown increasing invariance to nuisance factors. One particularly popular representation is based on the pictorial structures model [5, 6], where a template is associated to each body part. The articulation of one body part in relation to another in encoded as a relation orientation between the two. Often, a particular configuration of two or more body parts together is especially informative of the pose, and typically more easily detectable than the body parts independently. This idea gave rise to the introduction of poselets [7], templates that encode a specific configuration of body parts such as a bent arm. While poselets where initially used to detect humans, they have been employed for action detection from still images as well [8]. Such an approach works well for poses that are typical for a certain action, but it is less suitable for actions associated with arbitrary poses, without additional motion information.

Actions are not only characterizable by their pose or shape, but also by their movement over time. Jhuang *et al.* [11] show that movement at specific joint

locations gives strong cues for action recognition. Raptis and Sigal [9] include an optical flow term in their reformulation of poselets to recognize actions from video. In a related work by Yao *et al.* [10], optical flow is used to model transitions from one class of postures to the next. Effectively, this approach allows to detect actions that vary in their execution in time, such as interactions with a vending machine. There can be variation in the spatial performance of actions due to the environment. Reaching actions depend on the location of the object, and some have addressed this using object detections as cues for action recognition (e.g. [12, 13]). Actions can also have a social component if they are performed together with others. These actions are typically coordinated in the sense that movements of one person are affected by, and affect, the movements of others. In daily life, there are many coordinated interactions, such as walking hand-in-hand, dancing, shaking hands and fighting. Typically, the relative positions of people in the scene give rise to the understanding of interactions or group actions. Lan *et al.* [14] introduce the action context descriptor that encodes both the estimated action performed by the person under focus, and those in his vicinity. Choi *et al.* [15] address learning automatically the parameters of this vicinity in terms of size, distance and the division into discrete orientations. Besides people's relative positions, cues from people's orientation [16] and movements [17] can further help in understanding group activities. While much progress has been made in understanding these activities, the main focus is on broad action categories such as queueing or fighting. A notable exception is the work of Patron-Perez *et al.* [18], who focused on dyadic interactions such as hand shakes and hugs. However, they use upper body detectors and head pose classifiers without fine-grained limb movement information.

In these dyadic interactions, the movements of two people can be so tightly coordinated that it is intrinsically part of the interaction. For example, a hand shake would not be recognized if the two hands involved would not move in unison. Similarly, walking side-by-side and walking hand-in-hand are largely similar but differ in the coordination of the movements of the hands. Making a fine-grained distinction helps in understanding interactions, and the relations between people. From a detection perspective, it requires not only the detection of the actions of each individual involved in the interaction, but also the coordination of their movements. Often, there is a limited number of key joints or body parts involved. For a hand shake, these are typically the right hands of both interactants. In this paper, we introduce a method that builds on previous work in individual action recognition and extends it to detect fine-grained, coordinated interactions between two people.

3 Dyadic Interaction Detection

We present in this section our method to learn the poselet model to detect, in time and space, dyadic interactions from video. In the training phase, we also use Kinect information, to speed-up the learning of the poselets. A detection model is learned for each action individually. We discuss the training and evaluation of the models in the subsequent sections.

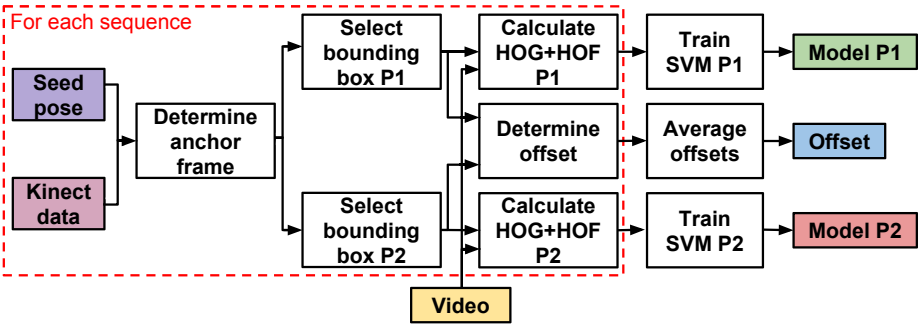

Fig. 1. Pipeline for learning the detection models for the two people involved in the interaction, and the spatial offset between the models. See text for details.

3.1 Training Action Models from Video and Kinect Data

In Figure 1 we show an overview of our pipeline for training the interaction models. We can identify four main parts in the pipeline which we will discuss one by one: first we determine the anchor frame to acquire the most invariant pose associated with the interaction, then we select the bounding box based on the joint configuration of the poselet to determine the best location to sample motion information from. After that we calculate the descriptors which, finally, are to be trained to acquire the detection model.

The first stage of the pipeline consists of locating the key joints of interest in the Kinect data. For a hand shake or a high five, these are the right shoulder, elbow, wrist and hand joints. For other interactions, other joints may be used. Using these key joints, we determine suitable frames in the learning data to extract a poselet that is representative for the particular interaction. Such an anchor frame can be regarded as containing a key pose. We hand-pick a seed frame at the epitome of the interaction from a random video. Next, we find in all other training sequences the most similar frame in terms of the Procrustus distance between the joint configuration of this seed frame and all frames in the sequence, following [7]. We rank all these frames based on the residual error. The frames below a certain threshold are ignored. Typically, we use a value between 0.5 and 0.75 as the threshold, which keeps the variance in the limb configuration low, while still retaining a sufficient number of sequences to learn the poselet from. The results of the selection procedure can be seen in Figure 2.

Based on this selection procedure, we retain a set of sequences to sample the frames from to create poselets for each of the two people involved in the interaction. As we know the anchor frame in each of the videos, we take this frame as the epitome of the interaction from which we sample the HOG descriptor. We consider the relative size and aspect ratio of all bounding boxes around the limbs involved in the interaction. These points determine two bounding boxes, one for each person. We warp the bounding boxes into the shape of the seed bounding box. The joint information from the Kinect metadata is also warped accordingly.

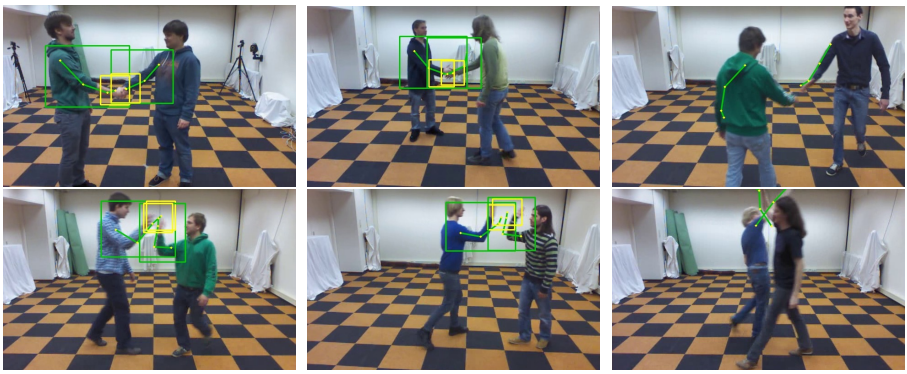

Fig. 2. Processed frames with respect to a seed frame (left): one with low (middle, both $RE = 0.15$) and one with high (right, $RE = 0.81$ and $RE = 0.52$) residual error. The bounding boxes are omitted in the rightmost images because they are not used for poselet creation, as the residual error exceeds the threshold.

Around these joints, we place a bounding box with a margin to the left top most joint (e.g., the right shoulder) and the right bottom most joint (e.g., the right hand). The margin size is chosen relative to the size of the skeleton in the seed frame.

After selecting suitable bounding boxes based on the joint locations provided by the metadata, we sample the HOG features for each of the two persons in the interaction. The size of the HOG template is determined by the size of the bounding box in the seed frame. The sizes of the two HOG templates are independent of each other. We use the HOG implementation described in [5] with both contrast insensitive (9 bins) and contrast sensitive features (18 bins), as well as 4 texture features. The resulting vector length is 31 data points per HOG cell.

While sampling the video data, we also track the relative positions of one person's limb with respect to the other's. This position allows us to calculate the mean offset of the limb bounding boxes. We process each of the sequences in our data in this manner. After extracting all the descriptors from the learning data, we determine the average offset of the two models with respect to each other. We also keep track of the relative positions of the joints within the poselet, to be able to sample the motion data from a specific position during the detection stage, without the need for additional pose information. The approach described here results in two poselets and their accompanying HOG representations. In Figure 3, we show the poselets for the hand shake interaction. These two HOG models are acquired by learning two separate linear Support Vector Machines (SVM), one for each class, that take the poselet vectors as positive data and randomly selected HOG vectors from the background images of the scene, as the negative learning data. Since we have a video stream from a static camera, it is easy to find frames that have little or no foreground information. Alternatively, random image data could be used. The models can also be trained using random

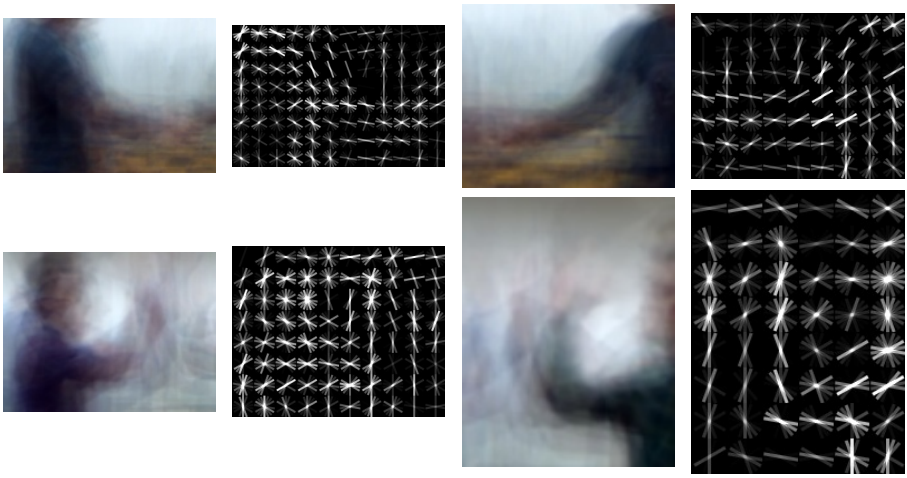

Fig. 3. Columns 1 and 3 show the normalized summed pixel values within the bounding boxes aligned by the anchor frame detection procedure, respectively for the right hand of the left and right person. Columns 2 and 3 show the corresponding HOG descriptors. Top row depicts the hand shake, the bottom a high five interaction. Note that the variance in the high five interaction is larger than in the hand shake interaction. This is due to the fact that the high five interaction is performed quicker, and has more variation in hand movement.

images that do not contain any humans, but we found that using the scenes background as negative learning data improves performance of the model. This is comparable to performing background subtraction.

Using the poselets, we can identify key poses of each of the two persons in the dyad. Since we are concerned with detecting fine-grained actions rather than just key poses, we enhance the HOG model with detailed movement information defined by a HOF descriptor. Adding motion information allows the detector to reject sequences where the poses match the model, though the motion information indicates otherwise. For instance, we can imagine two people standing hand-in-hand, being nearly indistinguishable from a hand shake by pose alone. We can find precise movement information within the poselet by taking advantage of the joint information provided by the Kinect. However, we want to create a model that will not depend on the Kinect data in the testing phase. This will increase the application potential of the method as video sequences are more commonly available. With the Kinect data in the training phase, not only can we speed up the selection of relevant frames, we can also learn a model that measures the movement of the actor only where it matters most: at the joint locations. It was shown in [11] that using the precise joint locations for movement measurements will give the best cues for action recognition. By defining a bounding box around the measured Kinect joint locations, we can enhance the HOG shape model with HOF movement descriptors, from which we classify the interaction more precisely.

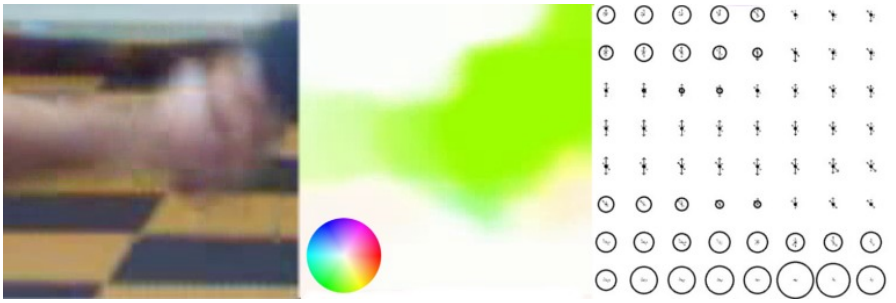

Fig. 4. Input image (left) with optical flow (middle) and visualization of the resulting HOF descriptor (right). The HOF visualization depicts directional arrows and circles. The length and direction of the arrows are indicative of the averaged magnitude and directional bin of the flow field histogram. The circles are indicative of the 0-bin of the histogram for parts of the flow field where no movement is present.

To build the HOF descriptors, we use an optical flow measurement inside a bounding box around the joints of the poselet. We create a HOF descriptor from the DeepFlow optical flow vectors [19] in the τ frames around the anchor frame. This sequence covers the movement at the joint location in such a way that a HOF descriptor with stable uniform movement directions can be created, as can be seen in Figure 4. The bounding box size is based on a square box on the seed frame. However, for the other sequences the bounding box may be warped by the same amount that the limb bounding box was warped due to the distortion of the limb, compared to the limb configuration in the seed frame. For the margin around the joint of interest on the seed frame we chose 32 pixels. The 15 joint motion sequence frames are clustered into 3×5 frames in which the movement directions are averaged for each of the 3 clusters. In each we create a directional grid much like the HOF descriptor described in [3]. However, a significant difference is that we use a regular grid instead of a flowing grid of keypoints, because we want our descriptor to be equally long for every measurement, as we do not use a bag-of-words approach for our classifier. The rightmost image of Figure 4 shows a visualization of our HOF output. Eventually, our detection model consists of a concatenation of this HOF descriptor to the HOG vector that describes the specific arm pose. The concatenated vector is learned by the soft margin linear SVM.

A separate SVM is trained for each of the two action classes. For negative learning data we use the scene's background for the static HOG part of the descriptor. For the HOF part of the negative feature descriptor we concatenate random movement samples from the dataset presented in [15].

3.2 Detecting and Classifying Interactions from Video Sequences

Given a sequence of frames, our method detects the occurence of a dyadic action in both time and space. For this, we use a sliding window approach. During testing, we use the two HOG/HOF models and the relative displacements of the

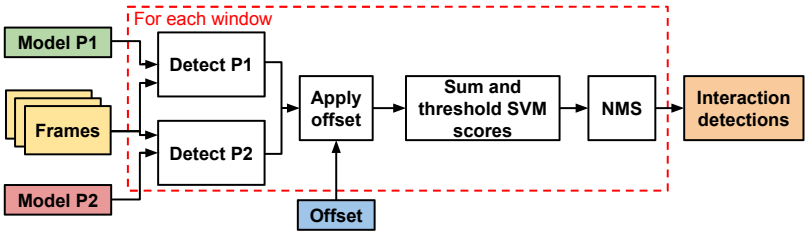

Fig. 5. Detection pipeline of the proposed approach. Refer to text for details.

models with respect to each other, see Figure 5. In the case of detecting hand shake and high five actions, we use the two models to detect the right arms of the two people involved in the interaction. We evaluate all possible locations of the poses and movements in a given frame. We apply the displacement offset and sum the results of the SVM scores on the detection results of the frame, shown in Figure 6. Then we use non-maximum suppression (NMS) to find the most likely position of the interaction of interest. The leftmost image in Figure 6 shows the detection of the right arm of the left person. There are false positive detections to the right of the right person. These scores are suppressed by adding the score of the right arm detection for the right person (second image). The sum detection result is shown in the third image. The fourth image shows the final detection after applying NMS. We show the ground truth location of the hand shake, determined by the Kinect, as a blue bounding box. The green bounding box is the top score that exceeds the threshold, the red bounding boxes are detections not exceeding the threshold.

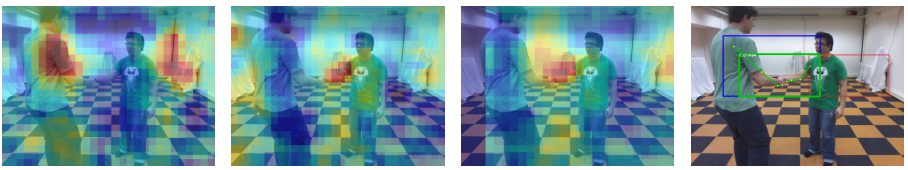

Fig. 6. Normalized detection scores for two single person models (first and second image), the cumulative score (third image) and the final detection result (fourth image)

4 Experiments and Results

In this section, we give an overview of our experiments and results. We first introduce our novel ShakeFive dataset[1], followed by a discussion of the baseline method, against which we compare our dyadic interaction detection method.

[1] The dataset is publicly available for research purposes and can be downloaded from `http://www.projects.science.uu.nl/shakefive/`

4.1 ShakeFive Dataset

There is a number of datasets available that focus on detecting human interactions from video. A well known interaction dataset is the TV Human Interaction Dataset [18]. It contains 300 video clips containing several interactions (hug, hand shake, high five and kiss) from TV shows. The clips are unconstrained in terms of viewpoint, lighting and occlussions. They are often cut mid action, which our approach cannot handle. The UT-Interaction dataset [20] contains six classes of continuous interactions between two people in 20 different videos. The interactions are recorded from a static viewpoint but the classes contain movements that are performed with different parts of the body.

To be able to perform experiments in our targeted setting with a fixed camera viewpoint observing fine-grained interactions, we have chosen to record our own ShakeFive dataset of dyadic interactions. The dataset consists of 100 RGB videos, as well as Kinect skeleton measurements for each individual involved. Each video contains 2 people who perform one of two possible interactions: a hand shake or a high five.

There are 57 videos containing hand shake interactions and 43 containing high five interactions. They are performed by a total of 37 unique individuals, 35 males and 2 females. The RGB video resolution is 640x480 pixels, recorded at 15 frames per second, with an average length of about 145 frames. In Figure 7, we show two example frames from the dataset. The videos are accompanied by metadata files that contain frame numbers, 20 skeleton joint positions per person acquired by Kinect (if there is a person in the frame), and one of 5 possible labels describing the interaction in the frame: standing, approaching, hand shake, high five, leaving.

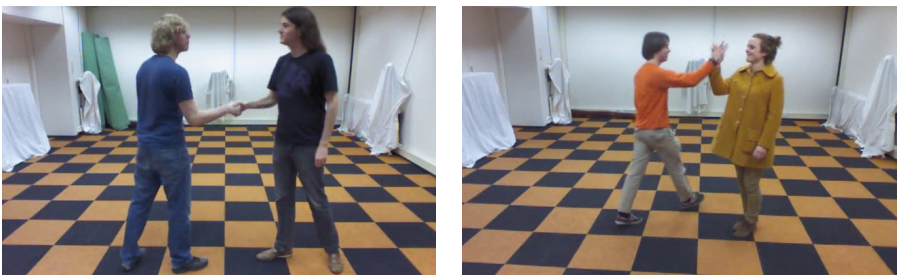

Fig. 7. Examples from ShakeFive dataset: hand shake (left) and high five (right)

4.2 Baseline

We compare our approach to the trajectory bag-of-words method of Wang *et al.* [3], which can be considered state-of-the-art. The method starts by finding trajectories of keypoints in an image sequence using a combination of HOG, HOF and Motion Boundary Histograms (MBH). Using these dense trajectories,

a bag-of-visual-words (BoVW) codebook with 4,000 clusters is created using k-means. As the dense trajectories are already normalized, we do not normalize any further during this step. Following [3], we then create the codebook from the best of 8 k-means clusterings, found by sampling 100,000 randomly chosen data points from the complete descriptor set. After creating the BoVW codebook, we use it to create a codeword for every 5 annotated frames in the training videos. Using these codewords, a binary classifier is learned using a Support Vector Machine. We use a Histogram Intersection Kernel, which has been shown to give superior results [21]. We normalize the codeword histograms using the L1-norm. During testing we generate an L1-normalized histogram codeword for the given frame, which is then fed to the classification model, leading to a binary classification which indicates whether or not the frame contains the interaction of interest.

In its original formulation, the baseline approach takes into account the entire image frame. Motion due to the camera or due to objects and other people in the foreground and background can affect the extracted trajectories in both the training and test sets. To make a closer comparison to our own implementation, we also run the baseline method using a sliding window approach in combination with NMS. We use the same codebook but take the average interaction window size to extract the keypoints to create the codeword at the particular sub-frame within the current frame. We train a new model on this input. During testing, we slide the window over the frame and classify every sub-frame, followed by NMS to obtain the final candidates, similar to the proposed method.

4.3 Results

The performance of the model is measured using a 4-fold cross validation, in which for the hand shake classification we use 42 randomly chosen video sequences, together with the additional hand picked seed frame video, making a total of 43 training videos containing a hand shake. We use the other 14 videos to test the model we train. The high five interaction is trained using 30 videos plus the seed video and is tested on the remaining 12 videos. For the sliding window approaches, both in the dyadic interaction detection and the baseline approach, we use a stride of 16 pixels in both directions. During testing, a window is counted as correct when more than 50% of it overlaps with the ground truth window in that frame.

We show the results of our experiments on the ShakeFive dataset in Table 1. We measure the performance of the algorithms with the Mean Average Precision (MAP) of the four folds. During testing we ran the detector on every 10^{th} frame of the video, from which we create short sequences of 15 frames ($\tau = 7$). This is approximately one second, and the value was empirically determined.

Our method scores an average precision of 49.56% using 75% of the available data for training and 25% for testing, using an exhaustive search with a sliding window on the tested frames. We have also tested the robustness of our method by reducing the amount of training data to 25%, while testing on the remaining three folds. This only slightly reduced the performance of the method both on

Table 1. MAP scores of the different methods on the ShakeFive dataset. For the proposed dyadic poselets method, we evaluated distributions of the training/test data of 75%/25% and 25%/75%.

Method	Hand shake	High five
Dyadic poselets (.75/.25)	**49.56**	**34.85**
Dyadic poselets (.25/.75)	47.87	23.94
Baseline	26.24	30.15
Baseline (sliding window)	20.10	23.07

the hand shake data and the high five data, showing its robustness. The hand shake data proves to be more robust than the high five data. Figure 8 makes the difference between the two types of interactions and the amounts of learning data clear. We can explain this by the speed of the interaction. A high five is a quick interaction compared to a hand shake. The sum images in Figure 3 show that the variation in the learning data is larger for the high five than it is for the hand shake. We can explain this by taking into account that the speed of the interaction in combination with the frame rate cause the matching of the poselets to be more difficult. This, in turn, makes the model less reliable, causing the larger drop in performance.

The baseline dense trajectory model gives an overall MAP performance for the implicit method for the hand shake action of 26.24% using the whole frame for classification and 20.10% for the sliding window approach. Here, the high five model performs slightly better (about 3-4%) than the hand shake model. We believe this is due to the fact that the dense trajectories on which the model is based are better capable of handling the fast motion of the high five. The performance of the baseline method stays below our method in both cases though. We believe this is because our model captures the pose and motion of the relevant body part for the given interaction more directly.

The sliding window approach on the baseline method slightly degrades performance on the baseline method. We believe this is due to the difference in the amount of classification windows. While the whole frame BoVW classifies a single interction per frame, using a sliding window approach in this case causes the system to classify more windows per frame. As a result the amount of false positives slightly increase because in considering more sub windows per frame, that do not contain the target action, the chance of a false detection increases.

Currently our method tries to find the frames that have limb configurations most similar to the limbs in the seed frame. This is not ideal. One of the issues here is that there are different view points of the interaction in the data, that will be missed by a model that is trained from one particular angle, which is bad for recall. Another issue is with the pose and motion slightly before and after the interaction. In the ground truth we chose to annotate the frame labels with some margin, so a hand shake for instance is already labeled as such when the person is moving his hand towards the hand of the other person. As we model

the handshake at the epitome of the interaction, the frames where the hands are not yet joined together will not be detected and marked as false negative. A temporal structure as proposed in [10] could be employed to solve this problem. Finally, we notice an issue with the placement of the joint locations over time. The update speed of the Kinect joints seems to be insufficient to follow the precise hand location during the interaction, causing the ground truth locations to be inaccurate at certain frames. While the Kinect does a good job at giving the vicinity of the limbs, we find that the final position of the end of the limbs, such as the feet or hands, is often inaccurate. In training, this in turn causes the location of where the motion is sampled from, for a given example with respect to the limb templatate in the frame, to be less than optimal.

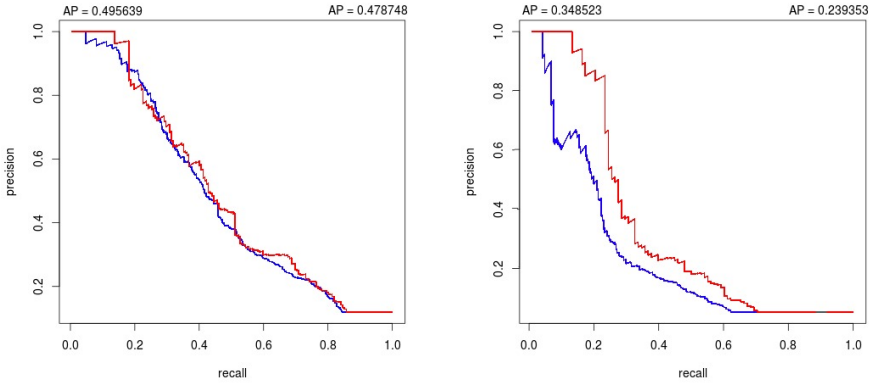

Fig. 8. The precision-recall diagrams for hand shake (left; red line, 0.75/0.25: $AP = 0.50$; blue line, 0.25/0.75: $AP = 0.48$) and high five (right; red line, 0.75/0.25: $AP = 0.35$; blue line, 0.25/0.75: $AP = 0.24$)

5 Conclusion and Future Work

We have presented an approach to detect, in space and time, fine-grained dyadic interactions from videos. We rely on a combination of pose and flow information to detect typical body poses of each person and the coordination of their joint movement, respectively. During training, we use joint pose information obtained using Kinect to speed up the selection of relevant frames to train the detectors. We have introduced the novel ShakeFive dataset that contains two-person hand shake and high five interactions to evaluate our method. In our experiments, our dyadic interaction detection method outperforms the baseline approach of Wang *et al.* [3]. Reducing the amount of training to 10-14 sequences only slightly reduces the performance in terms of average precision.

The motion information in the data provides us with good cues for enhancing the pose related to the coordinated interaction. We show how to add this motion information to a structured model that relies on poselets for interaction detection. The ShakeFive dataset contains the pose information that can be used as a stepping

stone for learning dyadic interactions. This is helpful in providing a simple way to get ground-truth information for the people involved in the interaction.

We identify several potential avenues for improvement of our approach. At this moment, we rely on a restricted viewpoint under which the the interaction is observed. During training, we ignore examples where the limbs of the people are too different from those in the seed frame. We would like to include these examples to achieve view invariancy to a certain extend. We can solve this by introducing multiple components, which model different angles separately, as in [5].

Currently, the contribution to our model from both HOG and HOF are weighted equally. In future work we would like to enhance the model and weigh the contribution of HOG and HOF depending on the motion and shape of the interaction. We intend to use the work by Mittal *et al.* [22] to perform structured output SVM ranking on the trained linear SVM outputs of our model.

We believe the placement of the joint locations from which the HOF is sampled could be improved by treating them as latent variables in our model. That means that the joint location in a poselet from which the motion is sampled, is not exactly fixed. This location acts as an initial indication of where the motion should be sampled from in the frame, but is then updated in an iterative fashion as is done in [5] for the location of the object parts with respect to the root of the template in that frame.

Finally, we will investigate to what extent the proposed approach can be used to detect dyadic, coordinated social behavior in conversations. The encouraging results on hand shakes and high five interactions have demonstrated the ability to detect fine-grained coordinated interactions involving the hands. We are interested to see how the method performs on interactions that involve other body parts, such as is the case with hugging and kissing. Ultimately, we would like to detect a range of meaningful dyadic interactions to better understand social behavior.

References

1. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
2. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: Proceedings International Conference on Pattern Recognition (ICPR), Cambridge, UK, pp. 32–36 (2004)
3. Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision (IJCV) 103(1), 60–79 (2013)
4. Felzenszwalb, P.F., Huttenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision (IJCV) 61(1), 55–79 (2005)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32(9), 1627–1645 (2010)
6. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 35(12), 2878–2890 (2013)

7. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: Proceedings IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, pp. 1365–1372 (2009)
8. Maji, S., Bourdev, L.D., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, CO, pp. 3177–3184 (2011)
9. Raptis, M., Sigal, L.: Poselet key-framing: A model for human activity recognition. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, pp. 2650–2657 (2013)
10. Yao, B.Z., Nie, B.X., Liu, Z., Zhu, S.C.: Animated pose templates for modeling and detecting human actions. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 36(3), 436–452 (2014)
11. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Proceedings IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, pp. 3192–3199 (2013)
12. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31(10), 1775–1789 (2009)
13. Yao, B., Fei-Fei, L.: Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 34(9), 1691–1703 (2012)
14. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(8), 1549–1562 (2012)
15. Choi, W., Savarese, S.: Understanding collective activities of people from videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 36(6), 1242–1257 (2014)
16. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of F-formations. In: Proceedings British Machine Vision Conference (BMVC), Dundee, United Kingdom, pp. 1–12 (2011)
17. Chang, M.C., Krahnstoever, N., Ge, W.: Probabilistic group-level motion analysis and scenario recognition. In: Proceedings IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, pp. 747–754 (2011)
18. Patron-Perez, A., Marszałek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in tv shows. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 34(12), 2441–2453 (2012)
19. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: Large displacement optical flow with deep matching. In: Proceedings IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, pp. 1385–1392 (2013)
20. Ryoo, M.S., Aggarwal, J.K.: UT-Interaction Dataset, ICPR contest on semantic description of human activities, SDHA (2010), `http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html`
21. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, pp. 1–8 (2008)
22. Mittal, A., Blaschko, M.B., Zisserman, A., Torr, P.H.S.: Taxonomic multi-class prediction and person layout using efficient structured ranking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 245–258. Springer, Heidelberg (2012)