Lend me a Hand: Auxiliary Image Data Helps Interaction Detection

Coert van Gemeren, Ronald Poppe and Remco C. Veltkamp
Interaction Technology Group, Department of Information and Computing Sciences,
Utrecht University, The Netherlands

***Abstract*— In social settings, people interact in close proximity. When analyzing such encounters from video, we are typically interested in distinguishing between a large number of different interactions. Here, we address training deformable part models (DPMs) for the detection of such interactions from video, in both space and time. When we consider a large number of interaction classes, we face two challenges. First, we need to distinguish between interactions that are visually more similar. Second, it becomes more difficult to obtain sufficient specific training examples for each interaction class. In this paper, we address both challenges and focus on the latter. Specifically, we introduce a method to train body part detectors from non-specific images with pose information. Such resources are widely available. We introduce a training scheme and an adapted DPM formulation to allow for the inclusion of this auxiliary data. We perform cross-dataset experiments to evaluate the generalization performance of our method. We demonstrate that our method can still achieve decent performance, from as few as five training examples.**

I. INTRODUCTION

The automated analysis of bodily behavior in video has seen impressive progress. Initial research considered behavior of individuals but the time seems right to proceed to the analysis of two people in close interaction. The way people interact with each other informs us of their activity, relation and the (cultural) context in which the interaction takes place. For example, when one person pats someone on the back, we can assume they are friends. Analyzing these interactions is therefore useful in areas such as social surveillance, video retrieval and human-robot interaction.

Human pose estimation and action recognition algorithms are increasingly robust to natural video data, partly due to the introduction of deep learning. Still, since algorithms rely strongly on the visual information, partial occlusions degrade the performance. In this paper, we look at two people in close interaction. The body parts that define the interaction are often occluded, which means we cannot rely on general algorithms for the estimation of human poses (e.g., [1], [2]).

Another challenge in recognizing interactions is dealing with differences in spatio-temporal coordination. Social interactions contain many semantically different but visually similar interactions. A handshake and passing an object are both described as two people facing each other with stretched arms. However, the coordination of the movement differs. Modeling spatio-temporal coordination in body pose and movement is therefore crucial. Despite an increasing focus on the inclusion of motion into human action detection models (e.g., [4]), it is rarely described at the level of body parts. This hinders modeling the fine-grained coordination of people's bodily behavior.

We focus on developing algorithms to detect large numbers of slightly different interaction classes in video. While the movement of an interaction is relatively stable across variations in environment settings, the pose can vary significantly as a result of differences in lighting, background and video quality. We depart from the trend of using increasing amounts of specific data to deal with these variations. For increasing numbers of classes, obtaining labeled data proves difficult. When the number of training videos is limited, we run the risk of over-fitting to their visual appearance. In this paper, we mitigate this risk by using images of unrelated actions and interactions. We investigate whether we can learn interaction detectors from a limited number of training videos per interaction. We learn movement cues from these videos, and use auxiliary data to ensure that pose cues generalize across environment settings.

Specifically, we train interaction detectors based on deformable part models (DPM, [5], [6]). We learn these models on the ShakeFive2 dataset [7], and test their generalization abilities on the UT-Interaction dataset [8]. We focus on three interactions: hand shake, passing an object and fist bump. Given their visual similarity, distinguishing between these interaction classes is challenging. To achieve generalization to environment settings, we use auxiliary images from the MPII Human Pose dataset [9]. This set contains images depicting humans involved in various actions, but none of the interactions that we consider.

Our contribution is two-fold. First, we introduce a framework to train deformable part models for interactions based on video data and auxiliary images. Second, we demonstrate increased generalization performance in cross-dataset experiments on spatio-temporal interaction detection.

We discuss recent advances in the analysis of interactions in the next section. In Section III, we introduce our interaction detection models, and detail how these can be trained from a limited number of videos and auxiliary images. We perform cross-dataset experiments and investigate the influence of the amount and type of training data in Section IV. We conclude in Section V.

II. RELATED WORK

Research on interaction detection is inspired by advances in individual action recognition [10]. Initial work has focused on local descriptors in a bag-of-visual-word approach. Wang et al. [11] track local descriptors over time to form dense

This work was supported by COMMIT and by NWO TOP-C2 grant 612.001.604 (ARBITER).

978-1-5090-4023-0/17/$31.00

trajectories that encode motion more robustly. Despite state-of-the-art action and interaction recognition results [11], [12], motion is not explicitly linked to specific body parts. Therefore, distinguishing between subtle interactions is difficult, especially when people interact in close proximity.

Instead of classifying directly from local features, recent work first estimates poses. Impressive results in 2D human pose estimation have been achieved using deep architectures such as convolutional neural networks (CNN) [2]. Knowledge of body pose has been shown to improve action recognition, and vice versa (e.g., [13]). In particular, the motion of the body improves action recognition performance [14], also in deep architectures [15].

While deep architectures work well for individuals, the presence of other people typically introduces occlusions. Because joints are estimated independently in a feed-foward manner, joints of different people can easily be confused when assembling full-body poses. This poses challenges for the detection of close proximity interactions where the coordination between people is essential. Several solutions have been explored to address this issue.

First, joint locations can be estimated conditionally, to encode pair-wise relations between joints. Gkioxari et al. [3] estimate joint locations using additional information from previously estimated joint locations. Insafutdinov et al. [1] estimate pair-wise joint relations directly from the image.

A second solution is to use stronger image cues. Poselets encode body parts in a specific articulation, such as a bent arm [16]. Such representations are less susceptible to occlusions, but the highly articulated nature of the human body requires a prohibitively large number of poselets for full-body pose estimation [17].

Finally, researchers have modeled the spatial relations between joints in a unified manner, in particular using deformable part models (DPM, e.g., [5], [6]). These are tree-like graphical models with body part templates as nodes and pair-wise spatial joint relations as vertices. Because the space of likely poses is limited for a given action, Iqbal et al. [18] condition both part templates and pair-wise relations on the action. Others have taken advantage of the more descriptive nature of poselets and have formulated DPMs with poselet parts [19], [20].

To further increase the modeling power of DPMs, researchers have included motion descriptors in the formulation. Yao et al. [21] describe actions as DPMs with a mixture of motion templates. Tian et al. [22] extended DPMs with spatio-temporal HOG3D descriptors. In fact, DPMs can be considered a specific type of CNN [23], which implies that handcrafted features in body part templates can be replaced by a learned representation.

Despite the powerful DPM formulation, occlusions during interactions remain an issue. To this end, Yang et al. [24] consider multiple people in a single DPM to model physical interactions such as a hand on someone's shoulder. Van Gemeren et al. [7] take this one step further by also modeling the characteristic motion for relevant body parts such as a lower arm in a hand shake.

While DPMs have much potential in the recognition and detection of interactions, they require sufficient numbers of positive training examples to determine the body part templates and pair-wise relations between parts. Especially when interactions are classified with more granularity, the amount of training data is typically low. This poses challenges in ensuring that the trained models generalize to videos with different visual properties. In this paper, we investigate whether we can learn the visual parts of DPMs independently from the motion parts and spatial relations between parts.

III. METHOD

We introduce the formulation of the interaction detection models in Section III-A. Based on an initially trained DPM, we adapt the body part templates using auxiliary data (Section III-B). Finally, we discuss how the trained DPMs can be used to detect interactions in video in both space and time.

A. Model Formulation

Within an interaction, we define the epitome as the moment where the pose and motion of two people are coordinated in a way that is characteristic for the interaction. We model both cues using a two-person DPM, as introduced by Van Gemeren et al. [7]. This interaction-specific DPM models both the pose and movement of two people in a single model. Each body part template describes the pose (histograms of oriented gradients; HOG), motion (histograms of flow; HOF) or both. Our templates resemble poselets as they encode body parts in specific orientations [16]. By including motion in the templates, we can also accurately describe the motion coordination between two people.

We formulate a DPM as graph $G = (V, E)$, with V a set of K body parts and E a set of connections between pairs of parts [6]. In this paper, we have $K = 5$ parts per person (torso, head, upper arm, lower arm and hand). Each part i $(1 \leq i \leq K)$ is centered on location $\boldsymbol{l}_i = (x_i, y_i)$. The scoring for a part configuration in image $\boldsymbol{I}$ is given by:

$$S(\boldsymbol{I}, \boldsymbol{l}) = \sum_{i \in V} \sum_{j \in D_i} \boldsymbol{w}_i^j \cdot \boldsymbol{\phi}_i^j(\boldsymbol{I}, \boldsymbol{l}_i) + \sum_{ij \in E} \boldsymbol{w}_{ij} \cdot \boldsymbol{\psi}(\boldsymbol{l}_i - \boldsymbol{l}_j) \quad (1)$$

The first term models the part appearance with a convolution of image feature vector $\phi_i(\boldsymbol{I}, \boldsymbol{l}_i)$ with trained detector $\boldsymbol{w}_i^j$. D_i is the set of feature representations for part i. In this paper, we consider HOG, HOF or a combination of both. However, our formulation is sufficiently flexible to incorporate a learned feature extractor such as CNNs [23]. The second term contains the pair-wise deformations between parts i and j, with $\boldsymbol{w}_{ij}$ the deformation cost of the connection.

We model interactions with a unified $2K + 1$-node graph with a virtual root node that branches to two sub-trees each representing the body parts of an actor. This part allows us to model relative distances between people as in [12], [25].

B. Training

We first learn a model for each interaction from a training set of positive examples and a generic source of hard negative data, in our case from [26]. Based on this initial model, we train the pose (HOG) parts of the detectors on auxiliary data and replace these parts in the detection model.

1) Learning an initial model: We start with a set of training videos for an interaction, each with estimated pose data. From these sequences, we select the most similar poses using a 2D version of the Kabsch algorithm [7]. We term the frames in which we find these poses the *epitome* frames. Based on the locations of selected joints (e.g., shoulder and elbow for the upper-arm), we train all body part templates individually using coordinate descent lSVM. For the HOF descriptors, we consider the movement in the 15 frames around the epitome.

We then assemble all body part templates into a single model. Each body part template's relative position is based on the average ground truth position of the body part over the epitome frames of all training videos. Deformation parameters, locations, biases and template features are optimized in a latent fashion by using the assembled initial model as a detector on the training examples. We harvest new latent positives as the best scoring detections. After this round of positive data optimization, we have a fully trained model that can be used for spatio-temporal interaction detection. Since a training set typically contains only a few positive examples, it is likely that our models overfit on these examples. To this end, we propose learning a more general model using auxiliary data.

2) Use of auxiliary data: In the auxiliary images, we intend to find right arms in a similar pose as the epitome frame in the interaction videos, both for the left and the right person. We select the images from the MPII Human Pose dataset that have a Kabsch distance for the right arm below a specified threshold κ. This results in two sets of images for the right arm: one for the left and one for the right person in the interaction. We train part templates for the upper-arm and lower-arm, for both the right and left person. We use the same feature representation and optimization method as for the original parts (see Section III-B).

We replace part templates $\boldsymbol{w}_i^j$ with the templates trained on auxiliary data $\hat{\boldsymbol{w}}_i^j$. Because the response of the body part templates might be different, the scoring of a part configuration (Equation 1) is also affected. Retraining the model on the initial training data to determine proper biases would undo the features of the replaced body part templates. Therefore, we transform the part scores in such a way that they fit the $0-1$ range.

Malisiewicz et al. [27] predict the overlap with the ground truth, given the response score of the model. Given a detection with HOG feature type of part i at location $\boldsymbol{l}_i$, the unary part response score US is:

$$US(\boldsymbol{I},\boldsymbol{l},i,j) = \frac{1}{1+e^{-\alpha_i(\boldsymbol{w}_i^j\cdot\boldsymbol{\phi}_i^j(\boldsymbol{I},\boldsymbol{l}_i)-\beta_i)}} \quad (2)$$

For HOF, the calculation remains the same and the original bias is used:

$$US(\boldsymbol{I},\boldsymbol{l},i,j) = \boldsymbol{w}_i^j \cdot \boldsymbol{\phi}_i^j(\boldsymbol{I},\boldsymbol{l}_i) \quad (3)$$

We can now replace the first term of Equation 1. The scoring function $\hat{S}$ using the replaced body part templates becomes:

$$\hat{S}(\boldsymbol{I},\boldsymbol{l}) = \sum_{i\in V}\sum_{j\in D_i} US(\boldsymbol{I},\boldsymbol{l},i,j) + \sum_{ij\in E} \boldsymbol{w}_{ij}\cdot\boldsymbol{\psi}(\boldsymbol{l}_i-\boldsymbol{l}_j) \quad (4)$$

We estimate parameters (α_i and β_i) as in [27], by measuring the ground truth overlap score of a detection with four randomly selected validation examples. This yields a sufficient number of detections to fit the sigmoid based on the part response score and the ground truth overlap.

C. Spatio-Temporal Localization

Trained models are used to perform spatio-temporal interaction detection. On a test video, we use a Gaussian pyramid to evaluate the models at different scales, for which the input resolution is halved every 10 layers. By applying the model, we obtain a series of detection scores (Equation 4).

We are looking at tubes of high-scoring detections but we only evaluate every 8^{th} frame. To obtain a continuous estimation of likely space-time locations, we map each detection to a 3D Gaussian [7]. In this 3D response space, we then find continuous tubes that cover an interaction from start to end in space and time. We start by localizing the highest score in the response space, and find the size of the closest detection. We then expand the tube forwards and backwards in time, centered on the response space and with the size of the closest detection. The expansion stops when neighboring frames do not sufficiently overlap, or when the score is below a threshold that is found during training.

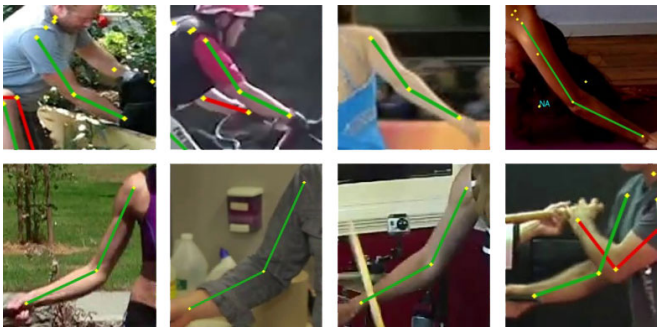


Fig. 1: MPII Human Pose images with similar arm pose. Matches are from different action categories. Note the variation in full body poses and visual appearance.

IV. EXPERIMENTS

We evaluate the performance of our interaction models trained both with and without auxiliary data. We consider the task of interaction *detection*: localizing interactions of a specific class in video in both time and space. This task is arguably harder than interaction *recognition* as we cannot rely on pre-segmented image sequences for classification. We focus on the recognition of hand shakes. To analyze how well

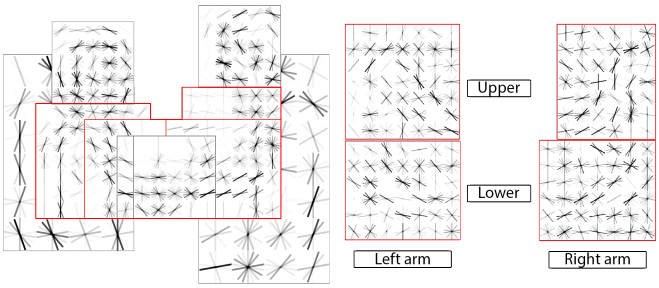


Fig. 2: Full model trained on ShakeFive2 (right) and arm parts trained from MPII (right). In red the respective positions (red) in the ShakeFive2 model.

we can distinguish between visually similar interactions, we also measure confusions with fist bump and object passing.

Because we are interested in generalization, we perform cross-dataset experiments. The datasets used for our experiments are described in Section IV-A. We describe detection metrics in Section IV-B. In Section IV-C, we discuss the setup of our evaluation. Results of these experiments are presented and discussed in Section IV-D.

A. Datasets

We use three publicly available datasets in our experiments. We train interaction models on the recently introduced *ShakeFive2* dataset[1] [7]. The *MPII Human Pose* dataset[2] [9] is used as a source of auxiliary data. We test on the *UT-Interaction* dataset[3] [8]. We briefly discuss these datasets.

Fig. 3: Example frames from ShakeFive2 and UT-Interaction.

ShakeFive2 consists of 94 videos of two people performing one of five close proximity interactions: *fist bump*, *hand shake*, *high five*, *hug* and *pass object*. In this paper, we only consider hand shake, fist bump and pass object as these are visually most similar. The interactions are recorded indoors under controlled settings but contain some variations in viewpoint (see Fig. 3(left)). For each person in each frame, 2D joint position data obtained using Kinect2 is available. Interactions are labeled per frame. For the spatial extent of an interaction, we use the minimum enclosing bounding box of both interactants.

[1] http://www.projects.science.uu.nl/shakefive/
[2] http://human-pose.mpi-inf.mpg.de/
[3] http://cvrc.ece.utexas.edu/SDHA2010/

UT-Interaction consists of videos of people performing close proximity interactions in two outdoor settings. The data is divided into two sets. The first set features at most two interacting persons at each moment, while the second set contains multiple pairs of people interacting simultaneously. Interactions *hand shake*, *hug*, *kick*, *point*, *punch* and *push* are performed. Pose data for these videos is not available but interactions are annotated per frame. As ground truth bounding boxes contain large margins and do not move with the people, we instead use the bounding box data from [25] as ground truth. Example frames appear in Figure 3(right).

MPII Human Pose contains close to 25k images with a total of over 40k people. The images are taken from Youtube and cover a broad range of daily human activities. There are 20 top-level action categories including dancing, music playing and water activities. None of the categories contains close proximity interactions. In particular, the data does not contain any frames with hand shakes. We ignore the action annotations and only use the pose data. Each person's 2D full body pose is annotated manually. Due to occlusions, there are missing joint locations. There is a significant amount of variation in viewpoint, lighting, clothing and image quality (see Figure 1 for an impression).

B. Performance Measurements

Our detection metrics reflect both the temporal and spatial accuracy as well as the label assigned to a detected interaction. We use the intersection over union (IoU) of the ground truth G and detected tube P as in [28]. G and P are sets of bounding boxes and θ is the set of frames in which either P or G is not empty. IoU of these two sets is calculated as:

$$IoU(G,P) = \frac{1}{\|\theta\|} \sum_{f \in \theta} \frac{G_f \cap P_f}{G_f \cup P_f} \tag{5}$$

To arrive at a single measure that considers recall and precision, we apply overlap threshold σ such that $IoU(G,P) \geq \sigma$ and report the Area under the Curve (AuC) averaged over all interactions in a test fold.

We are furthermore interested in the confusions between related classes. Therefore, we use the difference mean average precision (d-mAP) confusion matrix [7]. Each score in this matrix indicates how much of the AuC for a given class is lost to another class.

C. Experimental Setup

To test the performance of a model on an unsegmented test video, we find all candidate detections as described in Section III-C. The models have a temporal extent of 15 frames, and we process only every 8^{th} frame. Consequently, there is a temporal overlap of seven frames between subsequent candidates.

We consider two testing scenarios: *single class* (SC) and *multi-class* (MC). In the SC scenario, we test the spatio-temporal localization accuracy by applying the hand shake model on videos with hand shakes only. In the MC scenario, we additionally evaluate it on sequences with distractor classes fist bump and pass object. In this scenario, we test

how specific the detection model is. Confusions with the distractor classes will lead to lower AuC scores.

D. Results and Discussion

We first evaluate the interaction detection performance trained with and without auxiliary data. To highlight the generalization capabilities of our approach, we train on ShakeFive2 and test on UT-Interaction. To investigate how much training data is needed, we evaluate the impact of different amounts of auxiliary data and interaction videos on the detection performance.

We first train and test on the same dataset. By comparing the different models, we gain insight in the maximum performance on the same input data. HOGHOF is the model trained on ShakeFive2, and evaluated with Equation 1. Each body part template contains both HOG and HOF descriptors, except for the torso which only contains pose information. HOGHOF-SIG contains the same body part templates as HOGHOF ($\hat{\boldsymbol{w}}_i^j = \boldsymbol{w}_i^j$), but is evaluated using sigmoid scores (Equation 4). HOGHOF-AUX is the model trained with auxiliary data, evaluated using Equations 4. We search for similar poses with a Kabsch distance of $\kappa \leq 0.1$. For the left and right person respectively, we obtain 124 and 452 training images. In all tests, we use IoU overlap threshold $\sigma = 0.1$.

TABLE I: AuC for hand shake on ShakeFive2.

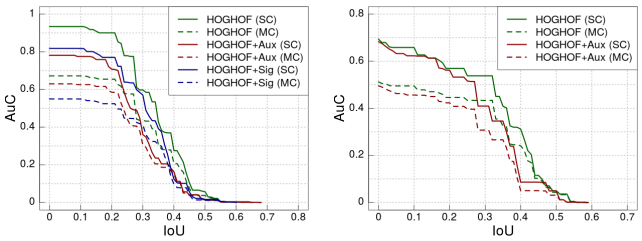
	SC	MC
HOGHOF	0.93	0.67
HOGHOF-SIG	0.82	0.55
HOGHOF-AUX	0.78	0.63

Table I summarizes the results on the detection of hand shakes on ShakeFive2. We use 4-fold cross-validation, with 5 training videos per fold. In the MC scenario, we additionally test on videos with fist bump and object passing interactions. Confusions with these classes results in a lower AuC. In both scenarios, there is a performance drop when using sigmoids. We attribute this to a lack of a bias that can give more weight to the detection score of specific body parts. With auxiliary data, the drop is partly compensated. We expect that this is due to the large number of positive examples, even though these originate from other data sources.

TABLE II: AuC for hand shake on UT-Interaction.

	Set 1		Set 2	
	SC	MC	SC	MC
HOGHOF	0.70	0.55	0.62	0.44
HOGHOF-AUX	0.74	0.52	0.50	0.39

In Table II the performance is shown for the detection of hand shakes on UT-Interaction. We train on all hand shake videos in ShakeFive2. Figure 4 shows the IoU graphs for both datasets. The drop in performance in ShakeFive2 when using auxiliary data (HOGHOF-AUX) instead of the original HOFHOF does not occur for UT-Interaction. Therefore, we believe the drop is due to overfitting on ShakeFive2.



(a) ShakeFive2 (b) UT-Interaction

Fig. 4: AuC scores for the hand shake in both MC and SC scenarios, on ShakeFive2 (left) and UT-Interaction (right).

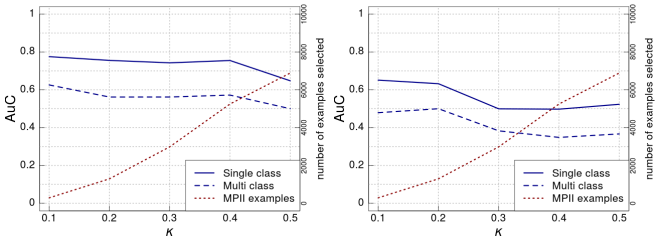
For the MC setting, we investigate confusions between interactions. In Table III we show d-mAP scores for ShakeFive2, without (left) and with (right) auxiliary data. Most confusions in HOGHOF occur for pass object, while hand shake is hardly confused with the other interactions. In general, slightly fewer confusions occur for HOGHOF-AUX than for HOGHOF.

TABLE III: D-mAP scores for HOGHOF (left) and HOGHOF-AUX (right) on ShakeFive2.

	FB	HS	PO
FB		0.30	0.34
HS	0.02		0.17
PO	0.20	0.30	

	FB	HS	PO
FB		0.21	0.26
HS	0.20		0.24
PO	0.08	0.16	

1) Varying Auxiliary Data: When selecting auxiliary data, there is a trade-off between the similarity of the pose and the number of training images. In Figure 5 we vary the Kabsch threshold κ. The performance on ShakeFive2 is hardly influenced by the degradation of the arm parts but more similar, but fewer, examples seems more suitable. On UT-Interaction, the effect is somewhat more pronounced. When κ increases from 0.1 to 0.5, the mean number of examples increases from 288 to 6886.

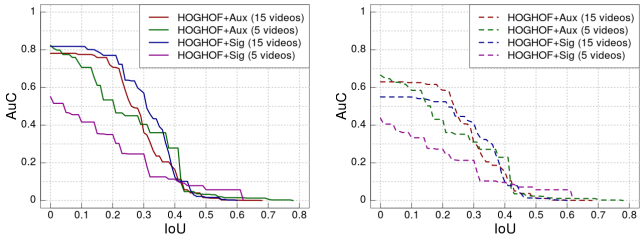


(a) ShakeFive2 (b) UT-Interaction

Fig. 5: AuC scores for the hand shake on ShakeFive2 (left) and UT-Interaction (right) for increasing κ.

2) Varying Training Data: We expect that using auxiliary data is more beneficial when fewer specific training examples are available. Previously, we have trained on 15 positive examples. In Figure 6, we show that this is indeed the case. When training on only five positive examples, the performance is largely retained while the lack of auxiliary data leads to a significant performance drop. This is promising as

auxiliary data is easily available.



(a) Single class (b) Multi-class

Fig. 6: AuC scores for the hand shake on ShakeFive2 with 5 vs. 15 training videos in both MC and SC scenarios.

V. CONCLUSIONS

We have introduced a novel method to train deformable part models for interactions using auxiliary data. By using body part templates learned from non-specific training data, we overcome overfitting due to the often limited number of available specific training examples. Our approach is especially useful in a cross-dataset setting.

Our deformable part models are trained on hand shake video examples from the ShakeFive2 dataset, with auxiliary images coming from MPII Human Pose. We apply these trained models on the UT-Interaction dataset. While performance is somewhat lower when the full number of training examples is used, we demonstrate that our novel approach performs better when the amount of specific training data is limited. We demonstrate decent hand shake interaction results with few confusions despite training on only five example sequences.

One limitation of our approach is that we rely on images with annotated poses. When we find relevant images based solely on the visual content, we could use even larger resources of auxiliary data. Similarly, we would like to explore whether auxiliary video examples could be used to better deal with variations in the movement in an interaction.

Such improvements should lead to a general framework to train detectors for large numbers of visually similar interactions. This will open up avenues to start addressing the understanding, rather than detection, of human interactions.

REFERENCES

[1] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proceedings European Conference on Computer Vision (ECCV) - Part 6*, 2016, pp. 34–50.

[2] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1653–1660.

[3] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *Proceedings European Conference on Computer Vision (ECCV) - Part 4*, 2016, pp. 728–743.

[4] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," in *CVPR - International Conference on Computer Vision and Pattern Recognition*, 2015.

[5] 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.

[6] 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.

[7] C. van Gemeren, R. Poppe, and R. C. Veltkamp, "Spatio-temporal detection of fine-grained dyadic human interactions," in *Proceedings Human Behavior Understanding Workshop (ECCV-HBU)*, 2016, pp. 116–133.

[8] 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.

[9] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3686–3693.

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

[11] 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.

[12] 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.

[13] B. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1293–1301.

[14] 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.

[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[16] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 1365–1372.

[17] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik, "Articulated pose estimation using discriminative armlet classifiers," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3342–3349.

[18] U. Iqbal, M. Garbade, and J. Gall, "Pose for action - action for pose," in *arXiv:1603.04037*, 2016.

[19] G. Gkioxari, B. Hariharan, R. B. Girshick, and J. Malik, "Using k-poselets for detecting people and localizing their keypoints," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3582–3589.

[20] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 588–595.

[21] 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.

[22] 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.

[23] 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.

[24] Y. Yang, S. Baker, A. Kannan, and D. Ramanan, "Recognizing proxemics in personal photos," in *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3522–3529.

[25] 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.

[26] 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.

[27] T. Malisiewicz, A. Gupta, and A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 89–96.

[28] 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.