DPM Configurations for Human Interaction Detection

Coert van Gemeren Ronald Poppe 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

Deformable parts models (DPMs) are a state-of-the-art models for the detection of human poses. Here, we investigate their performance on the detection of interactions between two people. We compare multi-layer to single-layer DPMs with either poselets or square parts placed on the limb joints of a skeleton configuration. The DPMs are based on popular models proposed by Felzenszwalb et al. [5] for generic object recognition in images, and the models proposed by Yang and Ramanan [16] for proxemics recognition, which is a method to recover postures of people interacting in photos. We train our human interaction models not on complete poses but only on the parts of the pose that are involved in the interaction. Our models are tested on five interaction classes[1] *from a novel human interaction data set:* ShakeFive2. *We show that poselets combined with multi-layered DPM yield the best results, increasing performance as much as 9.8% over a model with a single layer and 6.2% over a model without poselets.*

1. Introduction

Human interactions can be characterized by human poses. The poses people assume while engaged in an interaction give cues for assessing the interaction class [7]. For example, limb configurations when shaking hands are quite specific. In this work we are interested in detecting the spatially coordinated poses of two people simultaneously engaged in a specific interaction.

The automated estimation of human poses in images and video is challenging and has received a significant amount of research attention [10]. Previously proposed methods have encoded the pose implicitly, for instance using Bag-of-Visual-Features [14] or poselets [2, 8]. The latter are methods where areas covering parts of the body, while in a certain pose, are harvested, clustered and trained directly from the training data. In recent years, progress has been made using deformable part models (DPMs) [6], which encode a pose explicitly. There are some options in the configuration of the models. One can vary the number of parts and the resolution, size and location of each part. Another possibility DPMs offer is to replace the parts with poselets. In this work we compare DPMs with single or multiple layers, and with either parts or poselets.

We test the impact of choosing a multi-layered template model that has different resolutions for different parts, over a single-layered model that consists of a higher amount of smaller parts of the same resolution. We will also test models that combine different poselets in a DPM with either one or two layers. Our models are applied on a novel data set called *ShakeFive2*. It contains videos of two people performing several different interactions: *hand shake*, *high five*, *fist bump*, *pass object* and *thumbs up*. These actions all involve coordinated movement of the hands. The data set is challenging in the sense that the actions are visually similar. The video data is accompanied by metadata containing the skeletal configurations of the actors engaged in the interactions, obtained using [11]. This information is only used in training.

We make the following contributions. We introduce a data set containing several different interactions. We demonstrate robust detection of interactions using DPMs with various configurations. Furthermore, we show that a multi-layered model, consisting of parts of different resolutions is beneficial to the detection accuracy. Finally, we show that combining such multi-layered models with poselets as parts further increases the detection accuracy.

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

[1] Fist bump, hand shake, high five, pass object, thumbs up

2. Related Work

In this section, we review the literature on DPMs for the analysis of human poses and actions. DPM [5] has been employed for a variety of image analysis tasks, including pose estimation [16]. The original DPM is a two-layered template model using HOG cells, in which the bottom layer's cells have twice the spatial resolution of the cells in the top layer. When a part on the root layer consists of a template of $M \times N$ cells, a part on the other layer covering the same visual area is encoded by $2M \times 2N$ cells.

A variant of DPM was proposed by [16]. Though both models are based on the same deformable HOG templates, [5] considers a model with multiple layers that have parts of different resolutions. Furthermore, the locations of the fine-grained parts with respect to the root part are determined in an unsupervised manner. This approach is well suited for general object detection, such as in the Pascal Visual Object Classes (VOC) Challenge [3]. In contrast, the models in [16] have only one layer, that consists of multiple templates of $M \times N$ cells with the same resolution. The spatial relation between the parts provides important cues for the type of proxemic relation that is present in the visual data and is determined in a supervised manner. [16] manually annotate pairs of body parts from different people, that are physically in contact with each other during an interaction. The shape of this joint posture then defines their *proxemic relation*.

For the analysis of interactions between two people, both approaches may have some merit. The classic DPM can quickly find the general shape of an object through its root layer, after which the object detector refines the detection by checking whether it can find parts at approximately the correct locations within the root template. When searching an image for an interaction between two people, it is important to be able to quickly find basic human shapes. On the other hand, the most important cues for recognizing an interaction are provided by the limbs, for which there is a lot of variation in location and orientation. Limiting a model to a static root template will therefore not be suitable to recognize interesting poses. Yang and Ramanan [16] have considered personal photos of family and friends, in which the majority of the people are in the center of the image and face the camera. For the detection of interactions, this is typically not the case. How both these approaches deal with the specific challenges faced in the detection of coordinated interactions, is the focus of the current paper.

Poselets have been introduced with the aim of unconstrained person detection [1]. They are templates that are trained on a large amount of visual data of human subjects that has been clustered on the similarity of the poses. A poselet is not limited to a single body part; it may consist of several parts, as long as the spatial configuration of the parts that the pose is comprised of is consistently present throughout the visual data. The combination with DPM therefore seems promising. In this paper, we also investigate this combination by replacing the parts of [5, 16] with poselets.

3. DPMs For Interaction Detection

In this section we introduce the data set used in our research, and discuss the training and testing of the DPMs for human interaction detection, respectively.

3.1. *ShakeFive2* data set

The data set used in this research consists of 93 videos, each depicting one of five classes: *fist bump*, *hand shake*, *high five*, *pass object*, *thumbs up*. Every video in the data set was recorded using a static camera position, recording all the interactions from roughly the same view point. All videos are accompanied by a metadata file containing the action label and the joint locations of the actors at each frame. The skeleton data was obtained using [11], while the action labels were set by hand. Apart from the five interaction classes, each video has labels set before and after the actual interaction. These labels are either: *stand*, *approach* or *leave*. There are no unlabeled frames.

3.2. Training

To train a DPM, we define the body parts that we wish to learn, and the HOG cell's pixel resolution of each body part. A body part consists of a set of joints taken from the metadata. In Figure 1 we show how the body parts are sampled from the training data. We sample all defined body parts from both actors at the same time. We follow the method described by Van Gemeren et al. [13] to find the epitome frame for each interaction in each of the videos. The procedure to train a DPM using either root templates with parts, or poselets is essentailly the same. We anchor the parts or poselets to certain positions in the model, where they maintain a relative location with respect to each other.

When we have harvested the data for all parts that we want to learn from each video's epitome frame and each frame's skeleton, we create a mixture model for each opposing part. These have the same box color in Figure 1. The mixtures are optimized using the Dual Coordinate Descent SVM (DCD SVM) solvers presented in [4, 12]. After the positive optimization round, we perform a round of negative hard detection [5]. Negative hard examples are harvested in random frames of the Hannah data set used in [9], to avoid overfitting to the environment of the *ShakeFive2* data set. We avoid the locations containing people. After optimizing all part mixtures, we anchor all parts together. The anchor positions are determined by the relative offsets of the parts with respect to each other within the positive training data. This full DPM is then used on the positive training data to

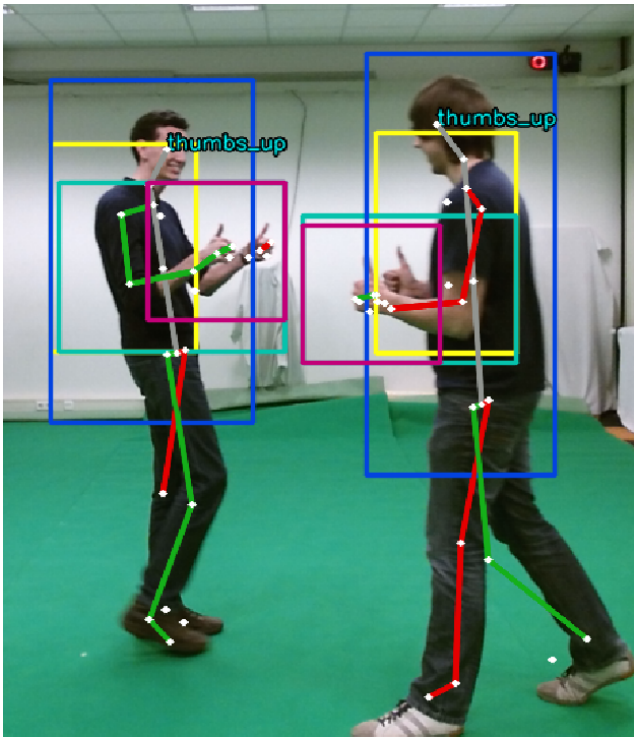

Figure 1. The bounding boxes of the body parts sampled from this *thumbs up* interaction frame each have their own colors: *torso* (blue), *right upper arm* (yellow), *right lower arm* (cyan), *right hand* (purple).

detect new latent positive examples. We consider all positive frames in this stage, looking for the highest scoring positive example in each training video to add to the positive examples set. Because we evaluate all positive frames during latent example harvesting, there is still a considerable amount of variation in the poses of the interactions, though only the pose that best fits the model in any particular video, is added as a new training example. The resulting positive examples are used for optimizing the model and to determine each part's bias and deformation parameters using the DCD SVM solvers.

There are two rounds of latent optimization at this stage. We have found that some parts tend to jump to incorrect positions during latent detection when no constraints are imposed. Therefore we measure the overlap of each part with the bounding box defined by the part's joint locations. We enforce a rule that only the detections that have a sum overlap for all parts on the frame, of more than 50%, are added to the positive examples set. In the second round of optimization this rule is ignored.

After optimizing the full model we test it on the fold of training videos that was left out. We do not test each frame of this video; rather we pick every eighth frame in the temporal region of positive interaction frames. Depending on the amount of positive frames picked from this region, we pick the same amount of frames before and after the interaction to test for false positive detections. At the start and the end of the interaction an onset and offset buffer of 16 frames is created from which no frames are picked. The exact start and end of the interaction tends to be somewhat ambiguous.

3.3. Testing

Detection is done using the DPM paradigm of [15]. The final detection is determined after performing non-maximum suppression (NMS) on all candidate detections of each frame. If the response score exceeds a threshold, the region is considered a detection.

A true positive detection overlaps for more than 50% with the ground truth bounding box, given by the hull created around all skeleton joints that were involved in creating the model. A detection outside of this ground truth bounding box, or a detection outside of the temporal positive interaction region, is considered a false positive.

4. Experiments and Results

We evaluate the described models for the detection of two-person interactions from video. We first focus on single-layer and multi-layer models, and then turn our attention to the combination of DPM and poselets.

4.1. Experiment Setup

In the first experiment, we compare multi-layer models with those with a single layer. We introduce four models with different part configurations. The standard multi-layer model (`ML`) has *torso* as a root part with a HOG cell size of 8 pixels, both horizontally and vertically. The torso template has a size of 4×7 cells. The other parts are anchored on the torso: *right shoulder*, *right elbow* and *right hand*. These parts are 5×5 cells in size with a resolution of 4 pixels per cell. We compare this model with a single-layer (`SL`) model in which the torso part is left out. The resolution and size of the other parts are the same as in the `ML` model. The `SL` model has a lower dimensionality than `ML`. To mitigate this, we also evaluate a `SL+head` model in which the head is added as an additional part. It has the same resolution and size as the other parts. Finally, we investigate whether the addition of the head part improves the performance of the `ML` model. To this end, `ML+head` is the `ML` model but with an additional high-resolution head part.

In the second experiment, we examine whether the use of poselets instead of parts improves the detection performance. We evaluate a single-layer and a multi-layer model and we use the best scoring model from the previous experiment as the baseline (`ML`). `P` is a single-layer model in which parts have been replaced by poselets. Specifically, the model contains the *right upper arm*, *right lower arm*

and *right hand*. The two opposing right upper arm poselets consist of the rectangular areas covered by the right shoulder joints and right elbow joints of the two actors in the frame, as can be seen in Figure 1. The right lower arm poselets consist of the areas covered by the right elbow joints and the right hand joints of both the actors. The resolution of the poselets is 4 pixels-per-cell. The multi-layer poselets model (`ML-P`) contains a root part (*torso*) and the three high-resolution poselets (*right upper arm*, *right lower arm* and *right hand*). Finally we test an unconstrained version of the `ML-P` model: `ML-PA`. Here we only perform one round of unconstrained latent positive detection and training. We add any detection to the positive examples set, that overlaps with the outer hull of the ground truth box for more than 50%, regardless of the part positions.

In both experiments, we perform 5-fold cross validation per interaction class. We measure the precision and recall for all five folds of the interaction class under consideration. Finally, we calculate the area under the curve as the average precision of the model. We have approximately 18 videos per interaction class, of which we have five.

4.2. Results For Experiment 1: Layers

Figure 2 shows that the model benefits from have multiple layers. It shows that the average precision for *Hand Shake* increases from 65% to 87% when we replace single-layer part templates (`SL`) by multi-layered torso templates with parts at half the resolution (`ML`). On average, it is clear that `ML` scores best over all interaction classes. Adding more parts to `SL` helps performance, increasing performance for *Fist Bump* from 39% for `SL` to 62% for `SL+head`. This is not always the case though. For *High Five* `SL+head` scores lower than `SL`. This is most likely due to the speed of the interaction. In case of a relatively fast moving interaction (e.g.: *High Five*), adding a part increases the noisiness of the model, compared to `SL`. This is not a problem for slow moving interactions, such as a *Fist Bump* or a *Hand Shake*.

In line with the findings for `SL+head`, `ML+head` does not seem to help performance much either, and in some cases (e.g.: *Hand Shake* and *Pass Object*) even decreases performance somewhat. The exception here is *Thumbs Up*, where `ML+head` scores slightly better than `ML`. This may be due to the fact that *Thumbs Up* is the only interaction we have tested where there is no physical contact between the two persons, which causes a larger spatial variation in the *torso* root templates. This problem is likely to be helped by adding a head part, which adds spatial freedom to the model by its deformation parameters. This added spatial freedom is also a likely cause of the slight decrease in performance for the other classes, where it is not needed and becomes a burden.

4.3. Results For Experiment 2: Poselets

In Figure 3 we compare the performance when we use `ML`, which was the best performing non-poselet model from the first experiment, to models that incorporate poselets. None of the poselet models outperform the multi-layer model, in fact the poselet only model `P` scores about 10% worse than `ML`. Clearly the poselets require some guidance from a rough estimation of the location of the torso.

The *Hand Shake* is the only class where `ML-P` outperforms the rest, scoring 3% better than `ML`. The biggest influence on the global AP score is the fact that `ML-P` scores 6% worse than `ML` in the *High Five* class. We believe the speed of the movement during the interaction is the main cause for the poselets to fail in this case. The relatively high speed of the right lower arm during a high five causes blurring of the video image, which translates to lower magnitudes in the orientation bins of the HOG cells. Another factor here is the fact that during a quick movement of a limb, their alignment over all positive examples is more difficult than during a slow movement, which causes more blurring of the gradients in the HOG cells during training.

When we examing *Thumbs Up* closely we notice that the unconstrained model (`ML-PA`) scores better than the constrained model (`ML-P`). This an odd result, because the constraint should improve performance by keeping bad latent examples out of the positive training set. The fact that `ML-P` scores worse might be because the *Thumbs Up* interactions takes place at a greater distance from one person to the other. It is the only interaction with no physical contact between the actors. This distance may cause the model to find only a few positive examples in the constrained case, because the variation in the part's locations is too large for the model to handle properly, resulting in a weak model. An unconstrained model on the other hand fixes the harvested part samples to whichever locations fit the model best, which may be the background in this particular case. This causes overfitting of the model, resulting in a better score than the constrained model.

5. Conclusions and Future Work

We have evaluated different DPM configurations for detecting, from an image, two people involved in a specific interaction. We have shown that DPM benefits of the combination of a multi-layered model with poselets. Contrary to [1], who also use a multi-layered approach, and [8], who use poselets to detect actions, we allow for a combination of poselets of different resolutions as in [5]. We have shown that the torso part provides the model with a good estimation of where the people are in the frame, and that this improves interaction detection performance. The detection is further refined by testing for the occurrence of the poselets in the frame, relative to the detected location of the torso.

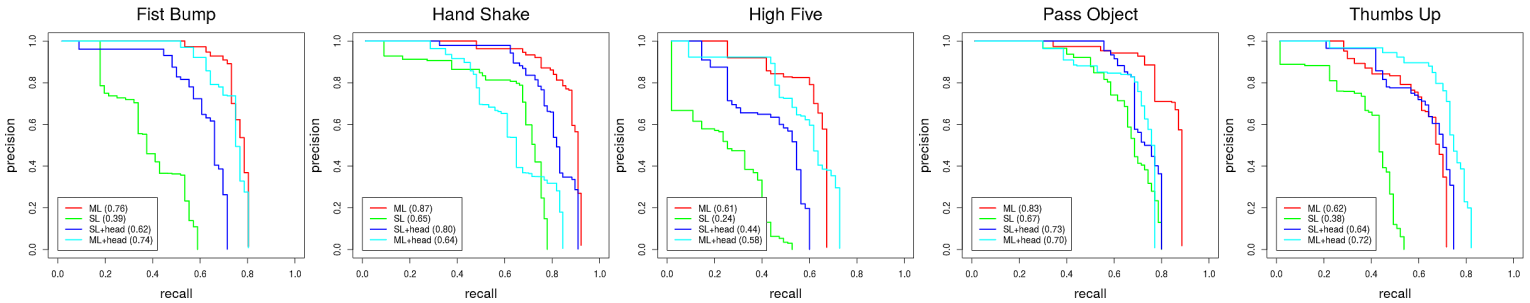

Figure 2. Model comparisons (best viewed in color). Average Precision (AP) mentioned in legend. Global averages over all classes mentioned after color names below.

Red (0.74) `ML`: Torso, Right Shoulder, Right Elbow, Right Hand.
Green (0.47) `SL`: Right Shoulder, Right Elbow, Right Hand.
Blue (0.65) `SL+head`: Head, Right Shoulder, Right Elbow, Right Hand.
Cyan (0.68) `ML+head`: Torso, Head, Right Shoulder, Right Elbow, Right Hand.

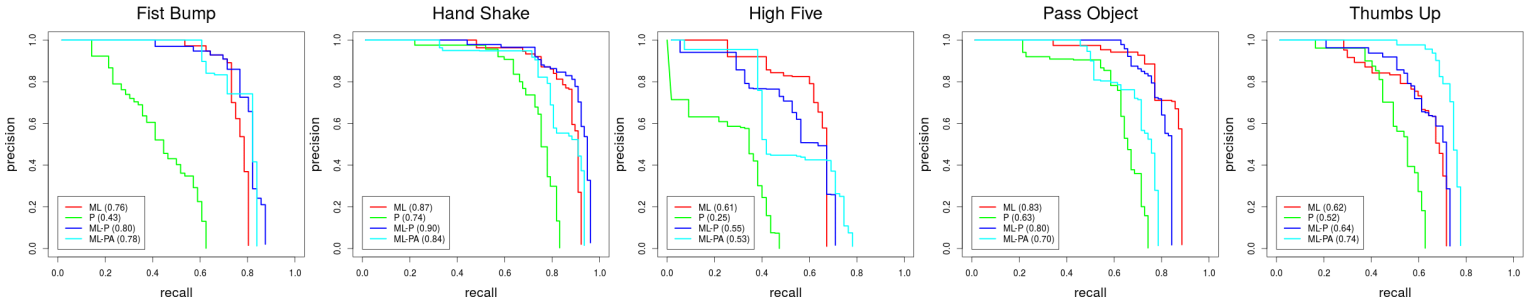

Figure 3. Model comparisons (best viewed in color). Average Precision (AP) mentioned in legend. Global averages over all classes mentioned after color names below.

Red (0.74) `ML`: Torso, Right Shoulder, Right Elbow, Right Hand.
Green (0.64) `P`: Right Upper Arm, Right Lower Arm, Right Hand.
Blue (0.74) `ML-P`: Torso, Right Upper Arm, Right Lower Arm, Right Hand.
Cyan (0.72) `ML-PA`: Torso, Right Upper Arm, Right Lower Arm, Right Hand. (1 round)

We have only considered detecting a specific interaction but are also interested in knowing how the different interaction models compete with each other in a recognition task. This requires that we test the models on all different classes. We assume that some classes are easily confused if the models are not trained discriminatively.

In this paper, we have considered a single viewpoint. Our work can be extended by adding more viewpoints at the top layer of the model. This way, we can improve detection performance when the poses vary a lot. Another extension of the models presented here would be to use parts that encode the movement of limbs rather than their shape. To create a model that can handle movement, we would have to include movement feature descriptors such as Histogram of Optical Flow (HOF). Such models would allow us to better discriminate between subtle, coordinated interactions that are visually similar such as shaking hands and passing an object.

References

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. *Proceedings ECCV 2010*, 2010.

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

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[4] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

[5] P. Felzenszwalb, R. Girshick, D. 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.

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 61(1):55–79, 2005.

[7] 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, Sydney, AUS, 2013.

[8] S. Maji, L. D. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3177–3184, Colorado Springs, USA, 2011.

[9] A. Ozerov, J. Vigouroux, L. Chevallier, and P. Pérez. On evaluating face tracks in movies. In *Proceedings IEEE International Conference on Image Processing (ICIP)*, pages 3003–3007, Melbourne, AUS, 2013.

[10] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, Washington, DC, USA, 2011.

[12] J. S. Supancic III and D. Ramanan. Self-paced learning for long-term tracking. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2379–2386, Portland, USA, 2013.

[13] C. Van Gemeren, R. T. Tan, R. Poppe, and R. C. Veltkamp. Dyadic interaction detection from pose and flow. In *Human Behavior Understanding*, pages 101–115, 2014.

[14] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (IJCV)*, 103(1):60–79, 2013.

[15] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, Colorado Springs, USA, 2011.

[16] 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.