Using Linking Features in Learning Non-parametric Part Models

Leonid Karlinsky and Shimon Ullman

Weizmann Institute of Science,
234 Herzl Street, Rehovot 76100, Israel
{leonid.karlinsky,shimon.ullman}@weizmann.ac.il
http://www.weizmann.ac.il

Abstract. We present an approach to the detection of parts of highly deformable objects, such as the human body. Instead of using kinematic constraints on relative angles used by most existing approaches for modeling part-to-part relations, we learn and use special observed 'linking' features that support particular pairwise part configurations. In addition to modeling the appearance of individual parts, the current approach adds modeling of the appearance of part-linking, which is shown to provide useful information. For example, configurations of the lower and upper arms are supported by observing corresponding appearances of the elbow or other relevant features. The proposed model combines the support from all the linking features observed in a test image to infer the most likely joint configuration of all the parts of interest. The approach is trained using images with annotated parts, but no a-priori known part connections or connection parameters are assumed, and the linking features are discovered automatically during training. We evaluate the performance of the proposed approach on two challenging human body parts detection datasets, and obtain performance comparable, and in some cases superior, to the state-of-the-art. In addition, the approach generality is shown by applying it without modification to part detection on datasets of animal parts and of facial fiducial points.

Keywords: linking features,parts detection,FLPM model.

1 Introduction

In this paper, we present a method for detecting parts of highly deformable objects. In these objects the parts configuration is flexible, and therefore the detection of the correct configuration is a difficult problem. As a central application we consider the detection of human body parts, namely: head, torso, lower and upper arms (figure 1). Recently, several approaches have been proposed to tackle this problem [1–10]. The central pipeline in all of these approaches can be roughly summarized as: input image → extraction of image features → estimating part posteriors using dedicated part detectors → extracting the most likely parts configuration by utilizing the human body kinematic constraints. For the last step of the pipeline prior parametric models of the human body are used,

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 326–339, 2012.

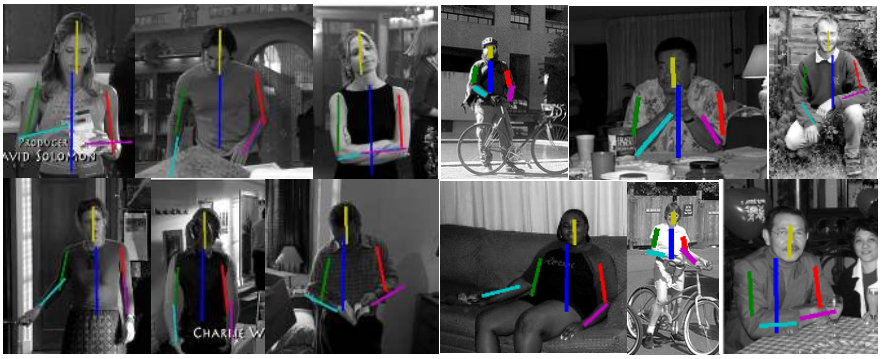


Fig. 1. Examples of detected body parts on the Buffy (left) and PASCAL stickmen (right). The detected parts are annotated using color-coded sticks.

which are based on the pictorial-structure (PS) model [1], with either standard tree-connectivity of the body parts [1–4, 8–10] (tree edges representing joints with constrained range of motion), or with extra non-tree edges for additional spatial constraints between physically unconnected parts [5–7, 11]. Notably, [10] use more parts in their tree PS model, allowing better treatment of non-rigid part deformations like foreshortening of limbs leading to better performance. In addition, [12] handle limb foreshortening in human pose estimation from video by employing stretchable models which model joint locations instead of limbs.

In the standard PS model, the pairwise constraints between the related parts are purely kinematic (e.g. a distribution of allowed relative angles) and do not depend on the query image. Recently, several approaches [8, 9] introduced some image dependence into the pairwise kinematic constraints of the PS model. In [8] the pairwise kinematic constraints are adaptive, and are derived from a subset of training images that are similar (in appearance) to the query image. In [9], connected parts are required to follow common segmentation boundaries, but only in the lower levels of the cascade, while in the higher cascade levels kinematic constraints are used. In addition, [13] uses the detections of poselet features to constrain the relative configuration between the parts. This state-of-the-art person detector [13] focuses on the detection of entire persons and their segmentation boundaries, and does not detect the arms and their sub-parts. Finally, a related idea of using chains of features to reach the part being detected has appeared in [14]; However, [14] detects only a single location on a part of interest (e.g. a hand), and its probabilistic model does not represent parts explicitly, leading to reduced detection performance compared with the proposed scheme (section 3).

In this paper, we propose a new method for parts detection that takes a different modeling approach. The approach follows the above pipeline of choosing the best configuration of detected part candidates. However, instead of estimating the quality of the relative configuration of part candidates using kinematic

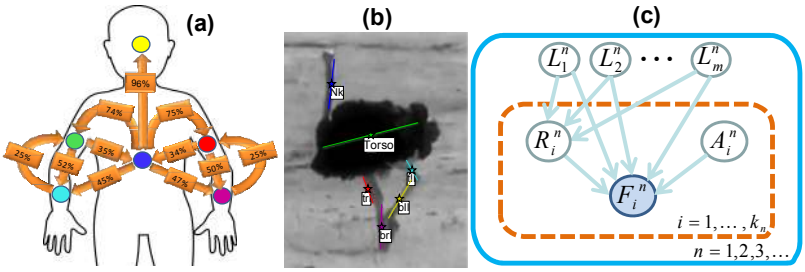


Fig. 2. (a) The learned associations between parts. An arrow from part A to part B shows how much part A affected (via the linking features) the detection of part B relative to all parts that affected part B. The incoming arrows sum to 100% (only arrows with at least 15% are shown); (b) Since no a-priori skeleton model is assumed (i.e. the object skeleton is not constrained by the model) the same approach can be directly applied to other objects (e.g. ostrich); (c) A graphical illustration of the FLPM model. The dotted orange and the solid blue rectangles are the features plate and the query images plate respectively. Only F_i^n variables are observed.

pairwise PS model constrains (either general or image-dependent), we propose a simple generative non-parametric Feature-Linked Parts Model (FLPM) that connects the part candidates through 'linking features' observed in the query image. Roughly speaking, the score of the relative configuration of two part candidates in a test image is measured by the cumulative score of the detected linking features. These features are patch descriptors learned, during training, to support this particular configuration. For example, to detect a particular bent-arm configuration, in addition to knowing that the relation between upper and lower arm parts is kinematically possible, it is useful to verify the appearance and location of features, such as the elbow, common to this arm configuration. In other words, the elbow appearance and possibly other image features common to this bent arm, 'link' the two parts of the arm together in this particular configuration.

As opposed to the kinematic models, no a-priori known model for the parts connectivity is assumed. The model implicitly adapts the parts connectivity pattern to each query image using the detected linking features associated with the different part-to-part connections. The adaptive nature of the connectivity pattern can also use image features that support relations between two kinematically unrelated parts. For example, crossed hands or hand over torso have some informative local image features associated with them that can be detected and used by the FLPM model. Figure 2a shows the average connectivity pattern obtained by averaging all of the linking-features based pairwise configuration scores from all the test images of the human body parts detection experiments (section 3).

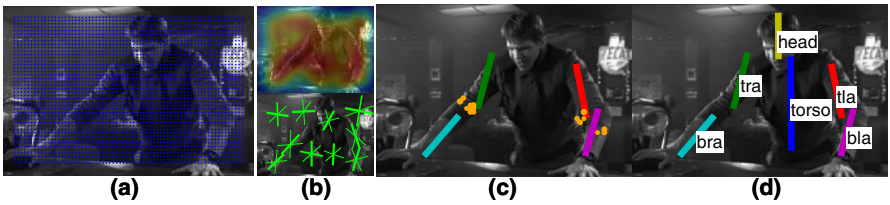


Fig. 3. Main stages of the proposed approach. (a) SIFT descriptor features are extracted on a regular grid (section 2.2); (b) Each part is voted for by the features (top shows a voting map for bottom left arm), and features are then clustered (section 2.4) to get candidates for the part (illustrated by lines on the bottom image); (c) Features vote for connections between parts and 'linking features' are identified to form the FLPM model for the test image, the example shows two automatically discovered sets of linking features (orange dots), for the lower-upper right arm and the lower-upper left arm connections respectively (section 2.1); (d) The final result is obtained through approximate inference on the obtained FLPM model (section 2.1). Best viewed in color.

The experimental results show that without known kinematic constraints, and using only simple SIFT descriptors computed on grayscale images as features (without more complex features such as color and segmentation contours used in [8, 9]) the approach achieves results comparable, and in some cases superior, to state-of-the-art on two challenging upper body datasets, Buffy [3] and PASCAL-stickmen [3]. In addition, since no a-priori known connectivity between the parts is assumed, the same approach can be successfully applied (without modification) to other domains. This is illustrated by applying the approach for facial fiducial points detection on the AR dataset [15, 16] (figure 5), and applying it for ostrich parts detection (figure 2b).

The rest of the paper is organized as follows. Section 2 provides the details of the proposed approach, section 3 describes the experimental evaluation; summary and discussion are provided in section 4.

2 Method

This section gives a brief overview of our approach further details are provided in subsections below.

The method starts with two standard pre-processing stages (figure 3a): (i) Extracting the region of interest (bounding box) around the object (e.g. human) whose parts are to be detected. It is a standard practice in pose estimation approaches, e.g. [3, 4, 8, 9], to assume that the object is already detected and roughly localized; and (ii) Computing image features (patch descriptors) centered on dense regular grid locations inside the region of interest. For the feature extraction stage, the step size of the grid and the size of the patches for which the descriptors are computed are determined by the scale of the region of interest, assuming the scale is roughly determined by the object detector. The details of the pre-processing are described in section 2.2.

After the pre-processing, the method proceeds in three main stages. Roughly, the first focuses on individual parts (figure 3b), the second on pairwise connectivity between part candidates (figure 3c), and the third on extracting the most likely global parts configuration (figure 3d). First, a set of candidate locations, orientations and sizes, are detected for each part. The detection of candidates is explained in section 2.4. Second, parameters of pairwise connectivity between part candidates are estimated, using non-parametric probability computations. At this stage, each configuration formed by a pair of part-candidates (e.g. relative position of two sticks), is scored by using the linking features (observed in the test image) that support this particular configuration. These linking features are patch descriptors that were associated with this configuration during training. The combined set of estimated connectivity parameters forms the FLPM model. Third, the most consistent subset of candidate parts is computed by a greedy approximate MAP inference over the estimated FLPM. The FLPM model, the computation of its parameters, and the approximate inference procedure for it are detailed in section 2.1.

The training phase of the method receives a set of images with annotated parts (e.g. stick annotation in upper body experiments). As the method is non-parametric, the training only involves building efficient data structures for similar descriptor search and memorizing a set of parameters for each feature, e.g. relative offset from different parts. These data structures are later used in order to compute the model probabilities for test images using Kernel Density Estimation (KDE) (section 2.3).

2.1 Feature-Linked Parts Model (FLPM)

Model Definition. The FLPM is a generative model which consists of the following random variables. $\left\{L_j^n\right\}_{j=1}^m$ are hidden variables describing the location, scale and orientation of the parts being detected in a test image I_n. In case the parts we are detecting are sticks (as in Buffy and PASCAL-stickmen datasets), the location is the (x, y) image location of the center of the stick, the scale is the length of the stick, and orientation is its $0 - 180$ degrees angle. We also add an additional part denoted 'full object', which is a union of all the parts, and its corresponding L_j^n has only location (of the center of mass of the object) and no orientation and scale. During the first step (section 2.4) we compute a discrete pool of candidates for each part and L_j^n takes its possible values from the corresponding pool for the $j-th$ part. $\{F_i^n\}_{i=1}^{k_n}$ are observed variables for the k_n features collected from the region of interest (around the object) on the test image I_n. In our implementation, F_i^n are the SIFT descriptors [17, 18] collected during the pre-processing stage (section 2.2) together with their image locations. The model includes two additional sets of auxiliary hidden random variables, which reside in the feature plate (i.e. two additional unobserved variables for each feature). These variables are: pair-of-parts choices $\{R_i^n\}_{i=1}^{k_n}$ and linking feature indicators $\{A_i^n\}_{i=1}^{k_n}$. The exact meaning of these additional variables is explained below the definition of the joint distribution. The FLPM model joint distribution is defined as:

$$P\left(\left\{L_j^n\right\}_{j=1}^m, \left\{F_i^n, R_i^n, A_i^n\right\}_{i=1}^{k_n}\right) = P\left(\left\{L_j^n\right\}_{j=1}^m\right) \cdot \prod_{i=1}^{k_n}\left[P\left(A_i^n\right) \cdot P\left(R_i^n \left| \left\{L_j\right\}_{j=1}^m\right.\right) \cdot P\left(F_i^n \left| A_i^n, R_i^n, \left\{L_j\right\}_{j=1}^m\right.\right)\right] \tag{1}$$

A graphical illustration of the model in plate notation is given in figure 2c.

The components of the model are explained below and the way to empirically estimate the necessary probabilities from the training data is explained in section 2.3. We do not impose any specific prior on the parts pose, therefore the joint $P\left(\left\{L_j^n\right\}_{j=1}^m\right)$ is assumed uniform. The A_i^n are binary hidden variables, one for each feature identifying the linking features in the image:

$$P\left(F_i^n \left| A_i^n \right., R_i^n, \left\{L_j^n\right\}_{j=1}^m\right) = \begin{cases} P\left(F_i^n \left| R_i^n, \left\{L_j^n\right\}_{j=1}^m\right.\right) & A_i^n = 1 \\ P\left(F_i^n\right) & A_i^n = 0 \end{cases} \tag{2}$$

Here $P\left(F_i^n\right)$ is the probability of spontaneously observing the feature F_i^n anywhere inside the region of interest regardless of the parts configuration. During the inference, the $A_i^n = 0$ option provides a more robust way of ignoring some of the features that are either outside the object or are not potential linking features. The R_i^n are pair-of-parts hidden variables taking values from the set of part pairs: $R_i^n = (p, q) \in \{(p, q) \,|1 \leq p \neq q \leq m\}$. The R_i^n provides the means to associate the feature F_i^n with a pair of parts from $\left\{L_j^n\right\}_{j=1}^m$. Roughly, $R_i^n = (p, q)$ means that the feature F_i^n connects parts p and q. The corresponding conditional is defined as follows:

$$P\left(F_i^n \left| R_i^n = (p, q), \left\{L_j^n\right\}_{j=1}^m\right.\right) = P\left(F_i^n \left| L_p^n \right., L_q^n\right) \tag{3}$$

The conditional probability of $R_i^n = (p, q)$ depends only on the assignemnts to L_p^n and L_q^n and is independent of the feature index i:

$$P\left(R_i^n = (p, q) \left| \left\{L_j^n\right\}_{j=1}^m\right.\right) = \eta_{pq}\left(L_p^n, L_q^n\right) \tag{4}$$

See section 2.3 for the way to compute $P\left(F_i^n \left| L_p^n \right., L_q^n\right)$ and $\eta_{pq}\left(L_p^n, L_q^n\right)$ from the training data. The η_{pq} can also contain any external prior on pairwise configuration of the parts, such as an articulation prior, and can be used to combine the proposed approach with other methods.

Inference. The MAP inference in the FLPM model involves the computation of the arg max, over part candidate choices $\left\{L_j^n\right\}_{j=1}^m$, of the following log-posterior:

$$\left\{L_j^{MAP}\right\}_{j=1}^m = \underset{\left\{L_j^n\right\}_{j=1}^m}{\arg\max}\left[\log P\left(\left\{L_j^n\right\}_{j=1}^m \left| \left\{F_i^n\right\}_{i=1}^{k_n}\right.\right)\right] \tag{5}$$

It is shown in Appendix A in the supplementary material that assuming that $P\left(A_i^n\right)$ is such that $P\left(A_i^n = 1\right) << P\left(A_i^n = 0\right)$, reflecting the fact that most of

the features are not potential linking features (or belong to the background), we can approximate the posterior as:

$$\begin{array}{l}\left\{L_j^{MAP}\right\}_{j=1}^{m} \sim \\ \underset{\left\{L_j^n\right\}_{j=1}^{m}}{\arg\max} \sum_{p,q=1}^{m} \left[\eta_{pq}\left(L_p^n, L_q^n\right) \cdot \sum_{i=1}^{k_n} \frac{P\left(F_i^n \middle| L_p^n, L_q^n\right)}{P\left(F_i^n\right)}\right]\end{array} \tag{6}$$

Note that although the posterior 6 does not depend on $\{A_i^n\}$, these auxiliary variables are necessary in order to derive it and make it more robust.

For a given test image I_n, the pre-processing (sec. 2.2) produces a set of k_n features $\{F_i^n\}_{i=1}^{k_n}$ and part candidates stage (sec. 2.4) produces a set of part candidates S_j for each of the m parts (such that for every j, $L_j^n \in S_j$). Assume for brevity that for every j: $|S_j| = r$, i.e. we have the same number of candidates for each part. In order to perform MAP inference, we compute the two tables W_R and W_F, both of size $m \times m \times r \times r$, each entry of which corresponds to a choice of $1 \leq p, q \leq m$ ($m \times m$ options), and choice of assignment to $\left\{L_p^n, L_q^n\right\}$ ($r \times r$ options). The $W_R\left(p, q, r_1, r_2\right) = \eta_{pq}\left(L_p^n = S_p\left\{r_1\right\}, L_q^n = S_q\left\{r_2\right\}\right)$ corresponds to pairs-of-parts prior, here $S_p\left\{r_1\right\}$ means the r_1-th element of the S_p set of part p candidates. The $W_F\left(p, q, r_1, r_2\right) = \sum_{i=1}^{k_n} \frac{P\left(F_i^n \middle| L_p^n = S_p\{r_1\}, L_q^n = S_q\{r_2\}\right)}{P\left(F_i^n\right)}$ corresponds to the feature based relative configuration prior. Note that the likelihood ratio inside the sum is higher for features that are more strongly associated with the considered relative part configuration. Finally, we multiply these two tables entry-wise (denoted $.*$) to form: $W = W_R .* W_F$ which is used for MAP inference over the FLPM posterior. According to equation 6, we need to find an optimal choice of indices $\left\{r_1^*, r_2^*, \ldots, r_m^*\right\}$ such that $\sum_{p=1}^{m} \sum_{q=1}^{m} W\left(p, q, r_p^*, r_q^*\right)$ is maximal. In our experiments, we have tested two greedy approximate inference approaches to the choice of these indices. In the first approach, denoted 'FLPM-sum', we start from the most likely candidate of the auxiliary 'full object' part and then greedily select parts and their optimal candidates having maximal sum of weights (entries in W) to the already selected parts:

$$q, r_q^* \leftarrow \underset{q, r_q}{\arg\max} \left[\sum_{p \in selected} W\left(p, q, r_p^*, r_q\right)\right] \tag{7}$$

In the second approach, denoted 'FLPM-max', the next part and its optimal candidate are greedily chosen using only the maximal weight to one of the already selected candidates:

$$q, r_q^* \leftarrow \underset{q, r_q}{\arg\max} \left[\max_{p \in selected} W\left(p, q, r_p^*, r_q\right)\right] \tag{8}$$

Experiments have shown that 'FLPM-sum' produces marginally better results than 'FLPM-max' and it was eventually the one we used for the FLPM inference.

This greedy inference procedure was selected based on its performance compared with alternatives. In particular, in our experiments, an optimization (for

maximizing $\sum_{p=1}^{m}\sum_{q=1}^{m} W\left(p,q,r_p^*,r_q^*\right)$) based on Quadratic Integer Programming (QIP) approximation produced results slightly inferior to the greedy scheme that we have eventually used.

2.2 Pre-processing and Feature Extraction

Given a training or test image I_n, and a bounding box around the object on that image, we extract the region of interest (ROI) by scaling the bounding box to a common 150 pixel height (in our implementation) and enlarging it by 20 pixels on every side. Exactly as in [3, 4, 8, 9], for the tested human body datasets the bounding boxes were automatically estimated from bounding boxes provided by [3]. The features $\{F_i^n\}_{i=1}^{k_n}$ extracted from the above ROI are SIFT descriptors computed for 20×20 pixel sized patches centered on a regular grid with 3 pixel step size.

The method parameters mentioned here and in subsequent sections 2.3 and 2.4, were set once (based on previous experiments with similar methods) and not further optimized; it is possible that optimization of the parameters for the current method could lead to even better perfromance.

2.3 Empirically Estimating FLPM Probabilities from the Training Images

The training phase of FLPM receives a set of images annotated with (body) parts on objects of interest. In our human upper-body experiments we used "sticks" annotation for parts [3], where each part is marked by a stick (two endpoints). No a-priori known connectivity pattern between the parts was assumed by our method. In order to obtain features associated with a specific part on each training image, we enlarge the part's stick in orthogonal directions to form a rectangle. Training features associated with the j-th part are collected from all the grid points (see section 2.2) residing inside the j-th part rectangles in all the training images. For each part, we construct an efficient Approximate Nearest Neighbors search data-structure (ANN) that allows efficiently searching for (approximate) neighbors of descriptors of test features among the descriptors of the training features. We use ANN implementation of [19] in our experiments. Denote by ANN_j the ANN containing the training features associated with the j-th part, and by ANN_{all} the ANN containing all the collected training features. Approximate Kernel Density Estimation (KDE) [20] over neighbors returned by querying the respective ANNs is used to compute all the feature related probabilities, namely: $P\left(F_i^n\right)$ - probability of F_i^n to appear spontaneously inside the object ROI; $P_j\left(F_i^n\right)$ - probability of F_i^n to appear somewhere on the j-th part (see section 2.4); $P_j\left(F_i^n \left| L_j^n \right.\right)$ - probability of F_i^n to appear in its particular location on j-th part (see section 2.4); and $P\left(F_i^n \left| L_j^n \right., L_k^n\right)$ - probability of F_i^n to "link" the j-th and the k-th parts.

Given a test feature descriptor F_i^n, the probabilities $P\left(F_i^n\right)$ and $P_j\left(F_i^n\right)$ are obtained by querying $q = 25$ neighbors of F_i^n in ANN_{all} and ANN_j respectively. The probabilities are then computed by KDE over these neighbors, as an

average (over the q neighbors) of $\exp\left(-0.5 \cdot d_r^2/\sigma^2\right)$, where d_r is the distance to the r-th neighbor ($1 \leq r \leq q$), and $\sigma = 0.2$. We assume that $P_j\left(L_j^n\right)$ and the $P\left(L_j^n, L_k^n\right)$ are uniform. Then the $P_j\left(F_i^n \left| L_j^n \right.\right) \propto P_j\left(F_i^n, L_j^n\right)$ is obtained using the neighbors of F_i^n returned by the ANN_j query. It is computed as an average of $\exp\left(-0.5 \cdot \left[d_r^2/\sigma^2 + \left\|o_j - o_r^j\right\|^2/\sigma_L^2\right]\right)$, where $\sigma_L = 15$ pixels, o_j is an offset between F_i^n's location (around which the descriptor was computed) and L_j^n's center location, and o_r^j is the offset between the location of the r-th neighbor and the center of the j-th part on the r-th neighbor source image. The $P\left(F_i^n \left| L_j^n, L_k^n \right.\right) \propto P\left(F_i^n, L_j^n, L_k^n\right)$ is computed as $0.5 \cdot \left(P_j\left(F_i^n, L_j^n, L_k^n\right) + P_k\left(F_i^n, L_j^n, L_k^n\right)\right)$, where $P_j\left(F_i^n, L_j^n, L_k^n\right)$ (and similarly $P_k\left(F_i^n, L_j^n, L_k^n\right)$) is obtained by the KDE over neighbors of F_i^n returned by the ANN_j query. The $P_j\left(F_i^n, L_j^n, L_k^n\right)$ is an average over r of:

$$\exp\left(-0.5 \cdot \left[d_r^2/\sigma^2 + \left(\left\|o_j - o_r^j\right\|^2 + \left\|o_k - o_r^k\right\|^2\right)/\sigma_L^2 + \left(\left\|\theta_j - \theta_r^j\right\|^2 + \left\|\theta_k - \theta_r^k\right\|^2\right)/\sigma_\theta^2\right]\right) \tag{9}$$

where d_r, o_j, o_k, o_r^j, and o_r^k are as above; θ_j and θ_k are orientations of the L_j^n and the L_k^n, and θ_r^j and θ_r^k are the orientations of the j-th and k-th parts respectively on the r-th neighbor source image; we use $\sigma_\theta = 20$ degrees. In addition, we restrict the above $P_j\left(F_i^n, L_j^n, L_k^n\right)$ computation only to neighbors which are identified as candidate 'linking features' during training. Those are the features, collected from training images, that lay in the intersections of the j-th and k-th part training rectangles enlarged by 10 pixels. This increases the importance of more relevant 'mutual' features for candidate parts consistency weighting (e.g. elbows for lower and upper arms).

Finally, we define $P\left(R_i^n = (p,q) \left| \left\{L_j^n\right\}_{j=1}^m \right.\right) = \eta_{pq}\left(L_p^n, L_q^n\right)$ to be uniform, thus providing no prior on the relative configuration of the parts. The $\eta_{pq}\left(L_p^n, L_q^n\right)$ can also incorporate the part candidate scores produced by the star model used to generate the candidates (section 2.4). However, in our experiments using part candidate scores did not result in performance increase due to the obvious difficulty of detecting parts independently of the other parts. In addition, any kinematic prior on the pairwise relative configuration of the parts that is used by the existing approaches may be incorporated into $P\left(R_i^n = (p,q) \left| \left\{L_j^n\right\}_{j=1}^m \right.\right)$ by substituting it with the prior value.

2.4 Generating Part Candidates

In this section we describe how the part candidates are generated for a given ROI in a test image I_n. The hidden variable L_j^n of the FLPM takes values from the candidates detected for the j-th part. The detection of candidates for the j-th part proceeds in three steps (same steps are repeated independently for each part to generate part candidates for all the parts).

First, all the collected features $\{F_i^n\}_{i=1}^{k_n}$ vote for the j-th part center of mass location using the non-parametric star model. This star model was used as one

of the baselines in [14] and is described in more detail there. Briefly, we accumulate votes from all the features in a matrix V_j with the size equal to the size of I_n, while each feature votes for 25 candidate locations of the part obtained from 25 approximate nearest neighbors, with voting weight equal to $P_j\left(F_i^n \left| L_j^n \right.\right) / P_j\left(F_i^n\right)$ (section 2.3). Each vote is smoothed by a Gaussian with STD of 15 pixels.

Second, up to $t \leq 100$ strongest local maxima $\{m_s\}_{s=1}^{t}$ are collected from V_j (with the weakest one m_t having at least 5% score of the V_j global maxima m_1: $m_t \geq 0.05 \cdot m_1$). Then for each feature F_i^n we compute a t-sized vector v_i each entry of which contains the amount of weight it contributed to each of the chosen maxima $\{m_s\}_{s=1}^{t}$. The features $\{F_i^n\}_{i=1}^{k_n}$ are then clustered into 20 clusters using spectral clustering [21], where the association between two features F_i^n and F_l^n is computed as an inner product between v_i and v_l. The reason for this clustering is that for the more ambiguous parts such as lower or upper arms, small local features along the boundary of the part have a wider range of possible part center locations relative to them (and they essentially vote for many of those).

Third, for each feature cluster the voting (as in the first step) is repeated using only features belonging to the cluster. The location of the maximal accumulated vote for each cluster is taken as the center location for the corresponding part candidate. Then the features of each cluster vote for the top three candidate orientations and top one length estimate of the part. The weights for orientation and length voting are the same as in the first step. This way, out of 20 feature clusters, 60 candidates are produced for each of the m parts.

After the computation of part candidates, the computation proceeds as explained in the description of the inference procedure in section 2.1. The next section describes the experimental evaluation of the proposed approach.

3 Results

To test the proposed approach we applied it to two challenging human body parts datasets proposed by [3], namely Buffy (version 2.1) and PASCAL stickmen (version 1.1). Moreover, in order to test the generality of the method, we also applied it (without modification) to the AR facial fiducial points detection dataset [15], and an ostrich parts detection dataset (see below). Table 4a summarizes the numerical results and state-of-the-art comparisons. The detections are assumed correct if they meet the standard $PCP_{0.5}$ criterion [3], where both endpoints of the detected part should be within 0.5 ground-truth part length from the ground-truth part endpoints (the full PCP curves are given in figure 5a). The results show that the FLPM produces comparable performance to the state-of-the-art approaches, improving the best result on the PASCAL stickmen v1.1 dataset. Note that FLPM is trained without knowing the kinematics of the object (connections and allowed relative angles between the parts) that is used by all the other methods [3, 8, 9], and using only grayscale images and simple SIFT descriptor features (as opposed to complex features involving color and segmentation boundaries used by [8, 9]). Figure 2a shows the average connectivity pattern discovered by the FLPM. This pattern is obtained by averaging all

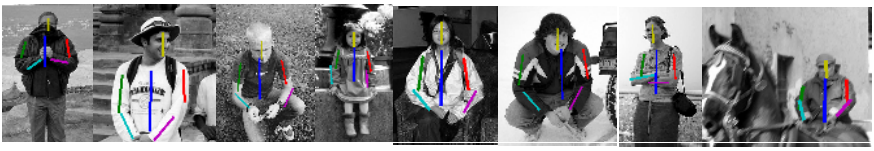
(a)

Buffy v2.1					
Method \ Part	**Torso**	**Upper arms**	**Lower arms**	**Head**	**Total**
FLPM	99.6%	93.2%	60.6%	99.6%	84.5%
candidate-recall	100%	98.1%	81.1%	100%	93.1%
Eichner & Ferrari [3]	98.7%	82.8%	59.8%	97.9%	80.1%
Andriluka et al. [4]	90.7%	79.3%	41.2%	95.5%	73.5%
Karlinsky et al. [14]	-	-	47%	-	-
Sapp et al. [8]	100%	91.1%	65.7%	100%	85.9%
Sapp et al. [9]	100%	95.3%	63.0%	96.2%	85.5%
Yang & Ramanan [10]	100%	99.6%	70.9%	99.6%	89.1%
Pascal Stickmen v1.1					
FLPM	98.8%	81.6%	47%	97.3%	75.5%
candidate-recall	99.8%	90.1%	71.9%	98.1%	86.9%
Eichner & Ferrari [3]	97.2%	73.8%	41.5%	88.6%	69.3%
Sapp et al. [9] (1)	99.3%	79%	49.3%	88.1%	74%
Yang & Ramanan [10] (2)	94.2%	79.7%	47.3%	93.1%	73.6%

(b)

Method \ Dataset	Buffy	PASCAL stickmen
FLPM	84.5%	75.5%
baseline: independent max score candidates (no linking features)	48.4%	33.9%

(c)



(1) both [15] and [16] originally reported their results on the (currently unavailable) PASCAL-stickmen v1.0 dataset containing less test images then v1.1, the result in the table is of applying the publicly available code of [16] on the v1.1 dataset.
(2) [10] did not report results for the PASCAL-stickmen dataset, the result in the table is of applying the publicly available code of [10] on the v1.1 dataset.

Fig. 4. (a) Summary of Buffy and PASCAL stickmen results. Our results are in blue. The 'candidate-recall' test computes the maximum recall rates (disregarding the score) of the individual part detectors (providing upper bound on the potential performance). To the best of our knowledge, [8] tested on the Buffy dataset version 1.0 (we test on v2.1); (b) Baseline comparison that tests the contribution of the core component of the model - the linking features. The comparison shows the prominent contribution of the linking features to the FLPM performance; (c) More examples of successfully detected parts on the tested datasets, illustrating a range of poses and appearances.

the entries of W (defined in section 2.1) corresponding to the part candidates selected during the inference stage for each of the test images. We also tested the maximal recall of the part candidates for the individual parts. In this experiment, denoted 'candidate-recall', we computed, for each part, the percentage of the test images on which at least one of the part candidates was correct ($PCP_{0.5}$). The results of 'candidate-recall' provide an upper-bound for FLPM performance and show that the correct part detection is included in the candidate set with high probability and therefore does not significantly limit the FLPM performance.

Table 4b shows FLPM vs. baseline comparison that tests the contribution of the core component of the FLPM – the linking features. Instead of choosing part candidates using the connectivity scores between the candidates (obtained by the linking features), for each test image, the baseline chose maximal scoring part candidate for each part. The candidate scores are produced by the star model that is used to generate them, as explained in section 2.4. Using these scores resulted in significant $> 35\%$ drop in performance, underscoring the importance of the linking features to the approach. The results of the baseline remained low even when adding a kinematic prior on the relative positions and angles

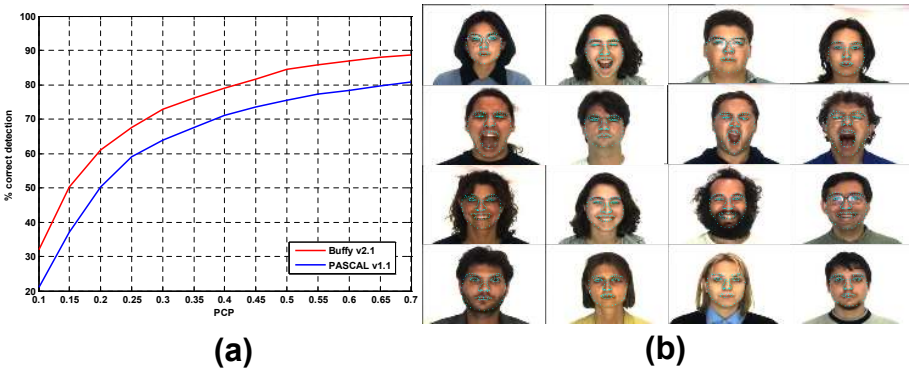


Fig. 5. (a) PCP curves for FLPM on PASCAL stickmen & Buffy datasets; (b) Example visual results of applying the FLPM approach (without modification) to the detection of 130 facial fiducial points (light blue) on the AR dataset.

between parts. The FLPM has two sources of information to select the most likely parts configuration, namely, the candidate part detection scores and the part connection scores via the linking features. The baseline illustrates that the most significant source of information in the model comes from the linking features. Figure 4c shows more examples of successfully detected parts.

Buffy v2.1. This dataset consists of 748 frames from 5 episodes (2-6) of 'Buffy the Vampire Slayer' TV series. On each image one person is annotated. The results are reported for the usual 235 images from the episodes 2, 5 & 6 detected by the upper body detector used in [3]. In our Buffy experiments, the training and testing is done in leave-one-episode-out manner, in which all images of one episode are used for testing, while images of all the other episodes are used for training.

PASCAL Stickmen v1.1. This dataset consists of 549 unrelated images with one annotated person per-image. In our experiments on this dataset, the training and testing was done in 5-fold cross validation manner, in which we divided the dataset images into 5 equal folds, and for testing on images of each fold the model was trained using images of the other folds. As in [3, 8, 9], the results are reported for the subset of 412 dataset images (denoted test images) on which the person bounding box was successfully detected by the upper body detector of [3]. Note however, that the two previous schemes [8, 9], originally reported their results on the previous (currently unavailable) version 1.0 of this dataset (with only 360 test images). Therefore, in table 4a, we compare the results of the proposed approach (FLPM) with the results of publicly available code of [9] applied to the 1.1 version of this dataset (412 test images).

AR Dataset. This is a facial fiducial points detection dataset proposed by [15, 16]. It consists of 895 images (with ground truth annotations) of 112 different people with 130 facial landmarks to be detected (figure 5b). The

images exhibit different facial expressions, different lighting conditions, and some occluders (glasses, facial hair, etc.). We treat each facial landmark as a part whose location is to be detected on the test images. The linking features in the FLPM model score the relative locations of the facial landmarks, and the most coherent (according to the FLPM) configuration of the facial landmarks is selected during inference. The experiments were conducted in leave-person-out-manner, each time leaving all the images belonging to a test person out of the training. The FLPM model attains an average detection error of 4.0 ± 1.3 pixels comparing favorably to the average error of 8.4 ± 1.2 pixels reported by the state-of-the-art method of [16].

Osterich. The dataset consists of 558 frames of a running ostrich movie sequence downloaded from YouTube. The training and testing was again done by 5-fold cross validation, each time leaving a sequence of about 110 consecutive frames for testing and training on the other frames. The average $PCP_{0.5}$ was 90.7%, with perfect detection of neck and torso, 78.2% detection of the upper legs and 93.8% detection of the lower legs. A movie with the visual results is provided in the supplementary.

4 Summary and Discussion

We presented an approach to the detection of highly flexible parts of deformable objects, which uses linking appearance features to weigh pairwise configurations between part candidates. Previous models have focused on the appearance of the individual parts combined with kinematic constraints (general or adaptive to the query image), the current model also learns and uses the appearance of part-linking. The results show that linking features supply highly useful information for identifying likely configurations of complex objects. In our implementation, linking features were used instead of the kinematic constraints that were used by other approaches. An interesting future research direction is to combine both the linking features constraints and the kinematic constraints in a single model, which can enjoy the power of both. For example, choosing the best result between the FLPM and the CPS method [9] for each test image of PASCAL stickmen v1.1, has the potential to improve the average correct detection to 81.4% (choosing all 6 parts together) or even to 83.7% (choosing each part separately). This suggests that the two methods are complementary on many test images and that a combined method is worth pursuing. In the current implementation, part candidates are discovered using dedicated star models, and are then connected by the linking features, which vote for consistent pairwise part configurations. One plausible future extension is to use a single model which will simultaneously discover parts and link them through intermediate linking features. A possible way to achieve this is to use an extension of the chains model [14], in which parts can be connected by extended chains of linking features (between parts) and internal features (within parts). Additional extensions include using more complex features and color information, which have been used successfully, e.g. in [8, 9, 13].

Acknowledgments. This work was supported by ERC Advanced Grant 269627 Digital Baby to SU. The authors would also like to thank Michael Dinerstein, Danny Harari, Maria Zontak, Nimrod Dorfman and the anonymous reviewers for valuable comments that helped improve this manuscript.

References

1. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV (2005)
2. Ramanan, D.: Learning to parse images of articulated objects. In: NIPS (2006)
3. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: BMVC (2009)
4. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR (2009)
5. Wang, Y., Mori, G.: Multiple Tree Models for Occlusion and Spatial Constraints in Human Pose Estimation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 710–724. Springer, Heidelberg (2008)
6. Jiang, H., Martin, D.R.: Global pose estimation using non-tree models. In: CVPR (2008)
7. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
8. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: CVPR (2010)
9. Sapp, B., Toshev, A., Taskar, B.: Cascaded Models for Articulated Pose Estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 406–420. Springer, Heidelberg (2010)
10. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: CVPR (2011)
11. Ioffe, S., Forsyth, D.: Human tracking with mixtures of trees. In: ICCV (2001)
12. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: CVPR (2011)
13. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV (2009)
14. Karlinsky, L., Dinerstein, M., Harari, D., Ullman, S.: The chains model for detecting parts by their context. In: CVPR (2010)
15. Martinez, A., Benavente, R.: The ar face database. CVC Technical Report num. 24 (1998)
16. Ding, L., Martinez, A.M.: Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. PAMI (2010)
17. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
18. Liu, C., Yuen, J., Torralba, A.: Sift flow: dense correspondence across different scenes and its applications. PAMI (2010)
19. Mount, D., Arya, S.: Ann: A library for approximate nearest neighbor searching. CGC (1997)
20. Duda, R., Hart, P.: Pattern classification and scene analysis. Wiley (1973)
21. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS (2001)