MODEC: Multimodal Decomposable Models for Human Pose Estimation

Ben Sapp
Google, Inc
bensapp@google.com

Ben Taskar
University of Washington
taskar@cs.washington.edu

Abstract

We propose a multimodal, decomposable model for articulated human pose estimation in monocular images. A typical approach to this problem is to use a linear structured model, which struggles to capture the wide range of appearance present in realistic, unconstrained images. In this paper, we instead propose a model of human pose that explicitly captures a variety of pose modes. Unlike other multimodal models, our approach includes both global and local pose cues and uses a convex objective and joint training for mode selection and pose estimation. We also employ a cascaded mode selection step which controls the trade-off between speed and accuracy, yielding a 5x speedup in inference and learning. Our model outperforms state-of-the-art approaches across the accuracy-speed trade-off curve for several pose datasets. This includes our newly-collected dataset of people in movies, FLIC, which contains an order of magnitude more labeled data for training and testing than existing datasets. The new dataset and code are available online. [1]

1. Introduction

Human pose estimation from 2D images holds great potential to assist in a wide range of applications—for example, semantic indexing of images and videos, action recognition, activity analysis, and human computer interaction. However, human pose estimation "in the wild" is an extremely challenging problem. It shares all of the difficulties of object detection, such as confounding background clutter, lighting, viewpoint, and scale, in addition to significant difficulties unique to human poses.

In this work, we focus explicitly on the multimodal nature of the 2D pose estimation problem. There are enormous appearance variations in images of humans, due to foreground and background color, texture, viewpoint, and body pose. The shape of body parts is further varied by clothing, relative scale variations, and articulation (causing foreshortening, self-occlusion and physically different body contours).

Most models developed to estimate human pose in these varied settings extend the basic linear pictorial structures model (PS) [9, 14, 4, 1, 19, 15]. In such models, part detectors are learned invariant to pose and appearance—*e.g.*, one forearm detector for all body types, clothing types, poses and foreshortenings. The focus has instead been directed towards improving features in hopes to better discriminate correct poses from incorrect ones. However, this comes at a price of considerable feature computation cost—[15], for example, requires computation of Pb contour detection and Normalized Cuts each of which takes minutes.

Recently there has been an explosion of successful work focused on increasing the number of modes in human pose models. The models in this line of work in general can be described as instantiations of a family of compositional, hierarchical pose models. Part modes at any level of granularity can capture different poses (*e.g.*, elbow crooked, lower arm sideways) and appearance (*e.g.*, thin arm, baggy pants). Also of crucial importance are details such as how models are trained, the computational demands of inference, and how modes are defined or discovered. Importantly, increasing the number of modes leads to a computational complexity at least linear and at worst exponential in the number of modes and parts. A key omission in recent multimodal models is efficient and joint inference and training.

In this paper, we present MODEC, a multimodal decomposable model with a focus on simplicity, speed and accuracy. We capture multimodality at the large granularity of half- and full-bodies as shown in Figure 1. We define modes via clustering human body joint configurations in a normalized image-coordinate space, but mode definitions could easily be extended to be a function of image appearance as well. Each mode is corresponds to a discriminative structured linear model. Thanks to the rich, multimodal nature of the model, we see performance improvements even with only computationally-cheap image gradient features. As a testament to the richness of our set of modes, learning a flat SVM classifier on HOG features and predicting the mean pose of the predicted mode at test time performs

[1] This research was conducted at the University of Pennsylvania.

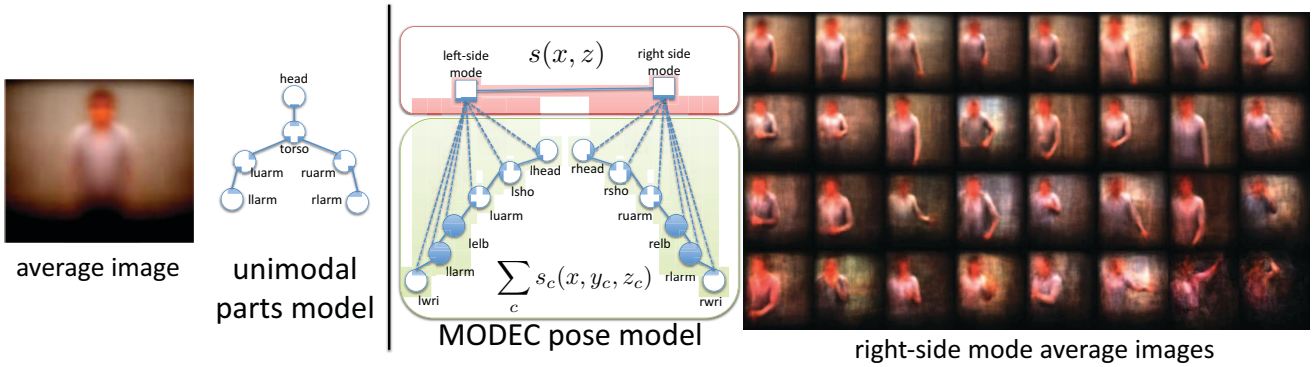

Figure 1. **Left:** Most pictorial structures researchers have put effort into better and larger feature spaces, in which they fit one linear model. The feature computation is expensive, and still fails at capturing the many appearance modes in real data. **Right:** We take a different approach. Rather than introduce an increasing number of features in hopes of high-dimensional linear separability, we model the non-linearities in simpler, lower dimensional feature spaces, using a collection of *locally* linear models.

competitively to state-of-the-art methods on public datasets.

Our MODEC model features explicit mode selection variables which are jointly inferred along with the best layout of body parts in the image. Unlike some previous work, our method is also trained jointly (thus avoiding difficulties calibrating different submodel outputs) and includes both large-scope and local part-level cues (thus allowing it to effectively predict which mode to use). Finally, we employ an initial structured cascade mode selection step which cheaply discards unlikely modes up front, yielding a $5\times$ speedup in inference and learning over considering all modes for every example. This makes our model slightly faster than state-of-the-art approaches (on average, 1.31 seconds vs. 1.76 seconds for [24]), while being significantly more accurate. It also suggests a way to scale up to even more modes as larger datasets become available.

Model	# part levels	# global modes	# part modes	training obj.
Basic PS	1	1	1	n/a
Wang & Mori [20]	1	3	1	greedy
Johns. & Evering. [11]	1	16	4*	indep.
Wang *et al.* [21]	4	1	5 to 20	approx.
Zhu & Ramanan [25]	1	18	1	indep.†
Duan *et al.* [3]	4	9	4 to 6	approx.
Tian *et al.* [18]	3	5	5 to 15	indep.
Sun & Savarese [17]	4	1	4	joint
Yang & Ramanan [24]	1	1	4 to 6	joint
MODEC (ours)	3	**32**	1	**joint**

* Part modes are not explicitly part of the state, but instead are maxed over to form a single detection.
† A variety of parameter sharing across global models is explored, thus there is some cross-mode sharing and learning.

Table 1. In the past few years, there have been many instantiations of the family of multimodal models. The models listed here and their attributes are described in the text.

2. Related work

As can be seen in Table 1, compositional, hierarchical modeling of pose has enjoyed a lot of attention recently. In general, works either consider only global modes, local modes, or several multimodal levels of parts.

Global modes but only local cues. Some approaches use tens of disjoint pose-mode models [11, 25, 20], which they enumerate at test time and take the highest scoring as a predictor. One characteristic of all such models in this category is that they only employ local part cues (*e.g.* wrist patch or lower arm patch), making it difficult to adequately represent and predict the best global mode. A second issue is that some sort of model calibration is required, so that scores of the different models are comparable. Methods such as [11, 13] calibrate the models post-hoc using cross-validation data.

Local modes. A second approach is to focus on modeling modes only at the part level, *e.g.* [24]. If n parts each use k modes, this effectively gives up to k^n different instantiations of modes for the complete model through mixing-and-matching part modes. Although combinatorially rich, this approach lacks the ability to reason about pose structure larger than a pair of parts at a time. This is due to the lack of global image cues and the inability of the representation to reason about larger structures. A second issue is that inference must consider a quadratic number of local mode combinations—*e.g.* for each of k wrist types, k elbow types must be considered, resulting in inference message passing that is k^2 larger than unimodal inference.

Additional part levels. A third category of models consider both global, local and intermediate part-granularity level modes [21, 17, 3, 18]. All levels use image cues, allowing models to effectively represent mode appearance at different granularities. The biggest downside to these richer models are their computational demands: First, quadratic mode inference is necessary, as with any local mode modeling. Second, inference grows linearly with the number of additional parts, and becomes intractable when part rela-

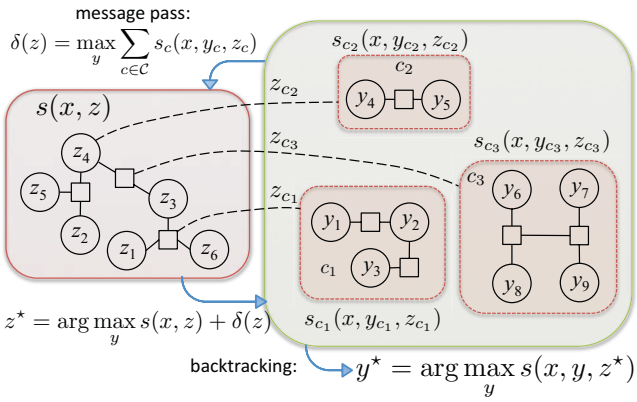

Figure 2. A general example of a MODEC model in factor graph form. Cliques z_c in $s(x, z)$ are associated with groups y_c. Each submodel $s_c(x, y_c, z_c)$ can be a typical graphical model over y_c for a fixed instantiation of z_c.

tions are cyclic, as in [21, 3].

Contrast with our model. In contrast to the above, our model supports multimodal reasoning at the global level, as in [11, 25, 20]. Unlike those, we explicitly reason about, represent cues for, and jointly learn to predict the correct global mode as well as location of parts. Unlike local mode models such as [24], we do not require quadratic part-mode inference and can reason about larger structures. Finally, unlike models with additional part levels, MODEC supports efficient, tractable, exact inference, detailed in Section 3. Furthermore, we can learn and apply a mode filtering step to reduce the number of modes considered for each test image, speeding up learning and inference by a factor of 5.

Other local modeling methods: In the machine learning literature, there is a vast array of multimodal methods for prediction. Unlike ours, some require parameter estimation at test time, *e.g.* local regression. Locally-learned distance function methods are more similar in spirit to our work: [10, 13] both learn distance functions per example. [12] proposes learning a blend of linear classifiers at each exemplar.

3. MODEC: Multimodal decomposable model

We first describe our general multimodal decomposable (MODEC) structured model, and then an effective special case for 2D human pose estimation. We consider problems modeling both input data x (*e.g.*, image pixels), output variables $y = [y_1, \ldots, y_P]$ (*e.g.*, the placement of P body parts in image coordinates), and special mode variables $z = [z_1, \ldots, z_K], z_i \in [1, M]$ which capture different modes of the input and output (*e.g.*, z corresponds to human joint configurations which might semantically be interpreted as modes such as *arm-folded, arm-raised, arm-down* as in Figure 1). The general MODEC model is expressed as a sum of two terms:

$$s(x, y, z) = s(x, z) + \sum_{c \in \mathcal{C}} s_c(x, y_c, z_c). \quad (1)$$

This scores a choice of output variables y and mode variables z in example x. The y_c denote subsets of y that are indexed by a single clique of z, denoted z_c. Each $s_c(x, y_c, z_c)$ can be thought of as a typical unimodal model over y_c, one model for each value of z_c. Hence, we refer to these terms as *mode-specific submodels*. The benefits of such a model over a non-multimodal one $s(x, y)$ is that different modeling behaviors can be captured by the different mode submodels. This introduces beneficial flexibility, especially when the underlying submodels are linear and the problem is inherently multimodal. The first term in Equation 1 can capture structured relationships between the mode variables and the observed data. We refer to this term as the *mode scoring term*.

Given such a scoring function, the goal is to determine the highest scoring value to output variables y and mode variables z given a test example x:

$$z^\star, y^\star = \arg\max_{z,y} s(x, y, z) \quad (2)$$

In order to make MODEC inference tractable, we need a few assumptions about our model: (1) It is efficient to compute $\max_y \sum_{c \in \mathcal{C}} s_c(x, y_c, z_c)$. This is the case in common structured scoring functions in which variable interactions form a tree, notably pictorial structures models for human parsing, and star or tree models for object detection, *e.g.* [8]. (2) It is efficient to compute $\max_z s(x, z)$, the first term of Equation 1. This is possible when the network of interactions in $s(x, z)$ has low treewidth. (3) There is a one-to-many relationship from cliques z_c to each variable in y: z_c can be used to index multiple y_i in different subsets y_c, but each y_i can only participate in factors with one z_c. This ensures during inference that the messages passed from the submodel terms to the mode-scoring term will maintain the decomposable structure of $s(x, z)$. A general, small example of a MODEC model can be seen in Figure 2.

When these conditions hold, we can solve Equation 2 efficiently, and even in parallel over possibilities of z_c, although the overall graph structure may be cyclic (as in Figure 1). The full inference involves computing $\delta(z) = \max_y \sum_{c \in \mathcal{C}} s_c(x, y_c, z_c)$ independently (*i.e.*, in parallel) for each z_c possibility, then computing the optimal $z^\star = \arg\max_z s(x, z) + \delta(z)$, then backtracking to retrieve the maximizing $y^\star$.

3.1. MODEC model for human pose estimation

We tailor MODEC for human pose estimation as follows (model structure is shown in Figure 1). We employ two mode variables, one for the left side of the body, one for

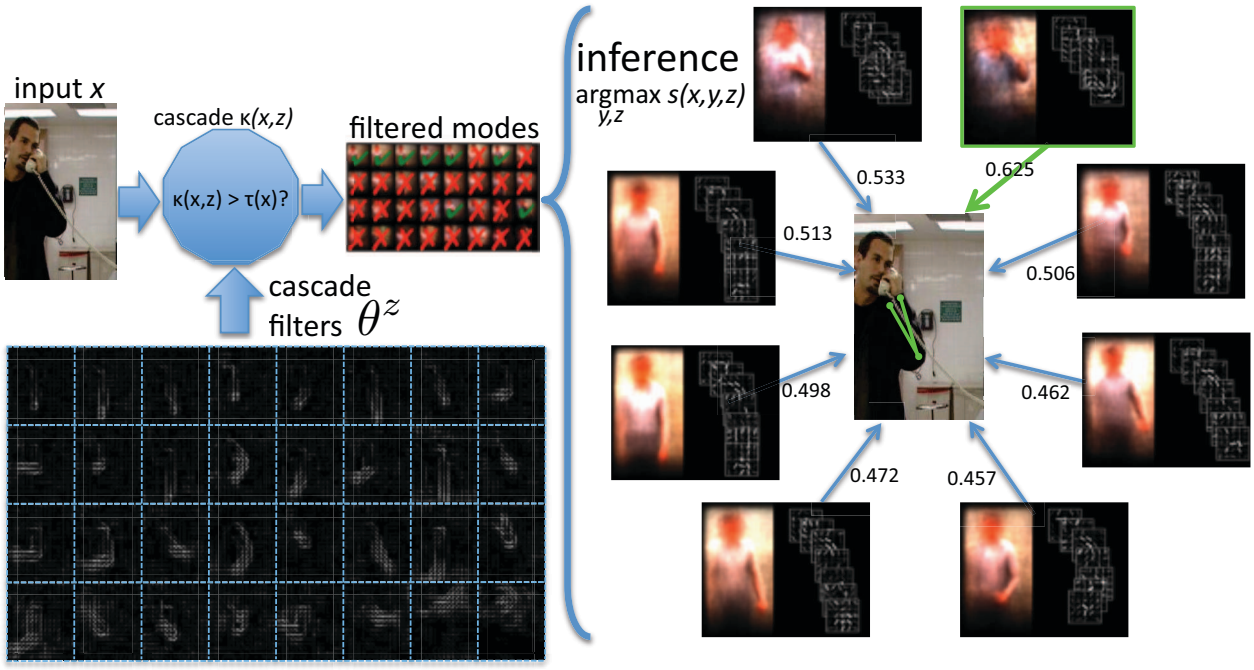

Figure 3. An illustration of the inference process. For simplicity, only a left-sided model is shown. First, modes are cheaply filtered via a cascaded prediction step. Then each remaining local submodel can be run in parallel on a test image, and the argmax prediction is taken as a guess. Thanks to joint inference and training objectives, all submodels are well calibrated with each other.

the right: z_ℓ and z_r. Each takes on one of $M = 32$ possible modes, which are defined in a data-driven way around clusters of human joint configurations (see Section 4). The left and right side models are standard linear pairwise CRFs indexed by mode:

$$s_\ell(x, y_\ell, z_\ell) = \sum_{i \in \mathcal{V}_\ell} \mathbf{w}_i^{z_\ell} \cdot \mathbf{f}_i(x, y_i, z_\ell) + \sum_{(i,j) \in \mathcal{E}_\ell} \mathbf{w}_{ij}^{z_\ell} \cdot \mathbf{f}_{ij}(y_i, y_j, z_\ell). \quad (3)$$

The right side model $s_r(\cdot, z_r)$ is analogous. Here each variable y_i denotes the pixel coordinates (row, column) of part i in image x. For "parts" we choose to model joints and their midpoints (*e.g.*, left wrist, left forearm, left elbow) which allows us fine-grained encoding of foreshortening and rotation, as is done in [24, 16]. The first terms in Equation 3 depend only on each part and the mode, and can be viewed as mode-specific part detectors—a separate set of parameters $\mathbf{w}_i^{z_\ell}$ is learned for each mode for each part. The second terms measure geometric compatibility between a pair of parts connected in the graph. Again, this is indexed by the mode and is thus mode-specific, imposed because different pose modes have different geometric characteristics. Details of the features are in Section 5. The graph over variable interactions $(\mathcal{V}_\ell, \mathcal{E}_\ell)$ forms a tree, making exact inference possible via max-sum message passing.

We employ the following form for our mode scoring term $s(x, z)$:

$$s(x, z) = \mathbf{w}^{\ell,r} \cdot \mathbf{f}(z_\ell, z_r) + \mathbf{w}^{\ell} \cdot \mathbf{f}(x, z_\ell) + \mathbf{w}^{r} \cdot \mathbf{f}(x, z_r) \quad (4)$$

The first term represents a (z_ℓ, z_r) mode compatibility score that encodes how likely each of the M modes on one side of the body are to co-occur with each of the M modes on the other side—expressing an affinity for common poses such as arms folded, arms down together, and dislike of uncommon left-right pose combinations. The other two terms can be viewed as mode classifiers: each attempts to predict the correct left/right mode based on image features.

Putting together Equation 3 and Equation 4, the full MODEC model is

$$s(x, y, z) = s_\ell(x, y_\ell, z_\ell) + s_r(x, y_r, z_r) + \mathbf{w}^{\ell,r} \cdot \mathbf{f}(z_\ell, z_r) + \mathbf{w}^{\ell} \cdot \mathbf{f}(x, z_\ell) + \mathbf{w}^{r} \cdot \mathbf{f}(x, z_r) \quad (5)$$

The inference procedure is linear in M. In the next section we show a speedup using cascaded prediction to achieve inference sublinear in M.

3.2. Cascaded mode filtering

The use of structured prediction cascades has been a successful tool for drastically reducing state spaces in structured problems [15, 23]. Here we employ a simple multiclass cascade step to reduce the number of modes considered in MODEC. Quickly rejecting modes has very appeal-

ing properties: (1) it gives us an easy way to tradeoff accuracy versus speed, allowing us to achieve very fast state-of-the-art parsing. (2) It also makes training much cheaper, allowing us to develop and cross-validate our joint learning objective (Equation 10) effectively.

We use an unstructured cascade model where we filter each mode variable z_ℓ and z_r independently. We employ a linear cascade model of the form

$$\kappa(x, z) = \theta^z \cdot \phi(x, z) \tag{6}$$

whose purpose is to score the mode z in image x, in order to filter unlikely mode candidates. The features of the model are $\phi(x, z)$ which capture the pose mode as a whole instead of individual local parts, and the parameters of the model are a linear set of weights for each mode, θ^z. Following the cascade framework, we retain a set of mode possibilities $\bar{M} \subseteq [1, M]$ after applying the cascade model:

$$\bar{M} = \{z \mid \kappa(x, z) \geq \alpha \max_{z \in [1,M]} \kappa(x, z) + \frac{1-\alpha}{M} \sum_{z \in [1,M]} \kappa(x, z)\}$$

The metaparameter $\alpha \in [0, 1)$ is set via cross-validation and dictates how aggressively to prune—between pruning everything but the max-scoring mode to pruning everything below the mean score. For full details of structured prediction cascades, see [22].

Applying this cascade before running MODEC results in the inference task $z^\star, y^\star = \arg\max_{z \in \bar{M},\ y \in \mathcal{Y}} s(x, y, z)$ where $|\bar{M}|$ is considerably smaller than M. In practice it is on average 5 times smaller at no appreciable loss in accuracy, giving us a $5\times$ speedup.

4. Learning

During training, we have access to a training set of images with labeled poses $\mathcal{D} = \{(x^t, y^t)\}_{t=1}^T$. From this, we first derive mode labels z^t and then learn parameters of our model $s(x, y, z)$.

Mode definitions. Modes are obtained from the data by finding centers $\{\mu_i\}_{i=1}^M$ and example-mode membership sets $S = \{S_i\}_{i=1}^M$ in pose space that minimize reconstruction error under squared Euclidean distance:

$$S^\star = \arg\min_S \sum_{i=1}^M \sum_{t \in S_i} ||y^t - \mu_i||^2 \tag{7}$$

where μ_i is the Euclidean mean joint locations of the examples in mode cluster S_i. We approximately minimize this objective via k-means with 100 random restarts. We take the cluster membership as our supervised definition of mode membership in each training example, so that we augment the training set to be $\mathcal{D} = \{(x^t, y^t, z^t)\}$.

The mode memberships are shown as average images in Figure 1. Note that some of the modes are extremely difficult to describe at a local part level, such as arms severely foreshortened or crossed.

Learning formulation. We seek to learn to correctly identify the correct mode *and* location of parts in each example. Intuitively, for each example this gives us hard constraints of the form

$$s(x^t, y^t, z^t) - s(x^t, y', z^t) \geq 1, \forall y' \neq y^t \tag{8}$$

$$s(x^t, y^t, z^t) - s(x^t, y, z') \geq 1, \forall z' \neq z^t, \forall y \tag{9}$$

In words, Equation 8 states that the score of the true joint configuration for submodel z^t must be higher than z^t's score for any other (wrong) joint configuration in example t—the standard max-margin parsing constraint for a single structured model. Equation 9 states that the score of the true configuration for z^t must also be higher than *all* scores an incorrect submodel z' has on example t.

We consider all constraints from Equation 8 and Equation 9, and add slack variables to handle non-separability. Combined with regularization on the parameters, we get a convex, large-margin structured learning objective jointly over all M local models

$$\min_{\{\mathbf{w}^z\}, \{\xi_t\}} \frac{1}{2} \sum_{z=1}^M ||\mathbf{w}^z||_2^2 + \frac{C}{T} \sum_{t=1}^T \xi_t \tag{10}$$

subject to:

$$s(x^t, y^t, z^t) - s(x^t, y', z^t) \geq 1 - \xi_t \qquad \forall y' \neq y^t$$

$$s(x^t, y^t, z^t) - s(x^t, y, z') \geq 1 - \xi_t \qquad \forall z' \neq z^t \in \bar{M}^t, \forall y$$

Note the use of $\bar{M}^t$, the subset of modes unfiltered by our mode prediction cascade for each example. This is considerably faster than considering all M modes in each training example.

The number of constraints listed here is prohibitively large: in even a single image, the number of possible outputs is exponential in the number of parts. We use a cutting plane technique where we find the most violated constraint in every training example via structured inference (which can be done in one parallel step over all training examples). We then solve Equation 10 under the active set of constraints using the fast off-the-shelf QP solver liblinear [7]. Finally, we share all parameters between the left and right side, and at test time simply flip the image horizontally to compute local part and mode scores for the other side.

5. Features

Due to the flexibility of MODEC, we can get rich modeling power even from simple features and linear scoring terms.

Appearance. We employ only histogram of gradients (HOG) descriptors, using the implementation from [8]. For local part cues $\mathbf{f}_i(x, y_i, z)$ we use a 5×5 grid of HOG cells, with a cell size of 8×8 pixels. For left/right-side mode cues

$\mathbf{f}(x, z)$ we capture larger structure with a 9×9 grid and a cell size of 16×16 pixels. The cascade mode predictor uses the same features for $\phi(x, z)$ but with an aspect ratio dictated by the extent of detected upper bodies: a 17×15 grid of 8×8 cells. The linearity of our model allows us to evaluate all appearance terms densely in an efficient manner via convolution.

Pairwise part geometry. We use quadratic deformation cost features similar to those in [9], allowing us to use distance transforms for message passing:

$$\mathbf{f}_{ij}(y_i, y_j, z) = [(y_i(r) - y_j(r) - \mu^z_{ij}(r))^2; \quad (y_i(c) - y_j(c) - \mu^z_{ij}(c))^2] \tag{11}$$

where $(y_i(r), y_i(c))$ denote the pixel row and column represented by state y_i, and μ^z_{ij} is the mean displacement between parts i and j in mode z, estimated on training data. In order to make the deformation cue a convex, unimodal penalty (and thus computable with distance transforms), we need to ensure that the corresponding parameters on these features $\mathbf{w}^z_{ij}$ are positive. We enforce this by adding additional positivity constraints in our learning objective: $\mathbf{w}^z_{ij} \geq \epsilon$, for small ϵ strictly positive[2].

6. Experiments

We report results on standard upper body datasets Buffy and Pascal Stickmen [4], as well as a new dataset FLIC which is an order of magnitude larger, which we collected ourselves. Our code and the FLIC dataset are available at http://www.vision.grasp.upenn.edu/video.

6.1. Frames Labeled in Cinema (FLIC) dataset

Large datasets are crucial when we want to learn rich models of realistic pose[3]. The Buffy and Pascal Stickmen datasets contain only hundreds of examples for training pose estimation models. Other datasets exist with a few thousand images, but are lacking in certain ways. The H3D [2] and PASCAL VOC [6] datasets have thousands of images of people, but most are of insufficient resolution, significantly non-frontal or occluded. The UIUC Sports dataset [21] has 1299 images but consists of a skewed distribution of canonical sports poses, *e.g.* croquet, bike riding, badminton.

Due to these shortcomings, we collected a 5003 image dataset automatically from popular Hollywood movies, which we dub FLIC. The images were obtained by running a state-of-the-art person detector [2] on every tenth frame of 30 movies. People detected with high confidence (roughly 20K candidates) were then sent to the crowdsourcing marketplace Amazon Mechanical Turk to obtain groundtruth labeling. Each image was annotated by five Turkers for $0.01 each to label 10 upperbody joints. The median-of-five labeling was taken in each image to be robust to outlier annotation. Finally, images were rejected manually by us if the person was occluded or severely non-frontal. We set aside 20% (1016 images) of the data for testing.

6.2. Evaluation measure

There has been discrepancy regarding the widely reported Percentage of Correct Parts (PCP) test evaluation measure; see [5] for details. We use a measure of accuracy that looks at a whole range of matching criteria, similar to [24]: for any particular joint localization precision radius (measured in Euclidean pixel distance scaled so that the groundtruth torso is 100 pixels tall), we report the percentage of joints in the test set correct within the radius. For a test set of size N, radius r and particular joint i this is:

$$acc_i(r) = \frac{100}{N} \sum_{t=1}^{N} \mathbf{1}\left(\frac{100 \cdot ||y_i^{t\star} - y_i^t||_2}{||y^t_{lhip} - y^t_{rsho}||_2} \leq r \right)$$

where $y_i^{t\star}$ is our model's predicted i^{th} joint location on test example t. We report $acc_i(r)$ for a range of r resulting in a curve that spans both the very precise and very loose regimes of part localization.

We compare against several state-of-the-art models which provide publicly available code. The model of Yang & Ramanan [24] is multimodal at the level of local parts, and has no larger mode structure. We retrained their method on our larger FLIC training set which improved their model's performance across all three datasets. The model of Eichner *et al.* [4] is a basic unimodal PS model which iteratively reparses using color information. It has no training protocol, and was not retrained on FLIC. Finally, Sapp *et al.*'s CPS model [15] is also unimodal but terms are non-linear functions of a powerful set of features, some of which requiring significant computation time (Ncuts, gPb, color). This method is too costly (in terms of both memory and time) to retrain on the $10\times$ larger FLIC training dataset.

7. Results

The performance of all models are shown on FLIC, Buffy and Pascal datasets in Figure 4. MODEC outperforms the rest across the three datasets. We ascribe its success over [24] to (1) the flexibility of 32 global modes (2) large-granularity mode appearance terms and (3) the ability to train all mode models jointly. [5] and [15] are uniformly worse than the other models, most likely due to the lack of discriminative training and/or unimodal modeling.

We also compare to two simple prior pose baselines that perform surprisingly well. The "mean pose" baseline sim-

[2]It may still be the case that the constraints are not respected (due to slack variables), but this is rare. In the unlikely event that this occurs, we project the deformation parameters onto the feasible set: $\mathbf{w}^z_{ij} \leftarrow \max(\epsilon, \mathbf{w}^z_{ij})$.

[3]Increasing training set size from 500 to 4000 examples improves test accuracy from 32% to 42% wrist and elbow localization accuracy.

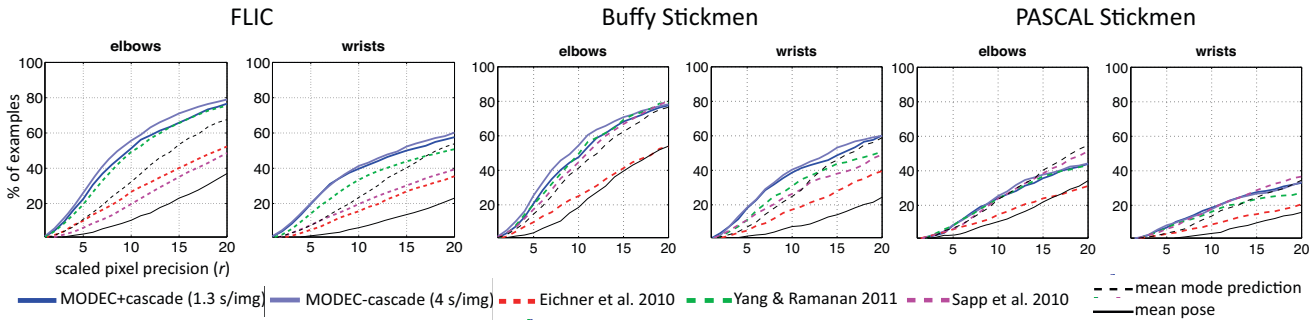

Figure 4. Test results. We show results for the most challenging parts: elbows and wrists. See text for discussion. Best viewed in color.

ply guesses the average training pose, and can be thought of as measuring how unvaried a dataset is. The "mean cluster prediction" involves predicting the mean pose defined by the most likely pose, where the most likely pose is determined directly from a 32-way SVM classifier using the same HOG features as our complete model. This baseline actually outperforms or is close to CPS and [5] on the three datasets, at very low computational cost—0.0145 seconds per image. This surprising result indicates the importance of multimodal modeling in even the simplest form.

Speed vs. Accuracy. In Figure 5 we examine the tradeoff between speed and accuracy. On the left, we compare different methods with a log time scale. The upper left corner is the most desirable operating point. MODEC is strictly better than other methods in the speed-accuracy space. On the right, we zoom in to investigate our cascaded MODEC approach. By tuning the aggressiveness (α) of the cascade, we can get a curve of test time speed-accuracy points. Note that "full training"—considering all modes in every training example—rather than "cascaded training"—just the ones selected by the cascade step—leads to roughly a 1.5% performance increase (at the cost of $5\times$ slower training, but equal test-time speed).

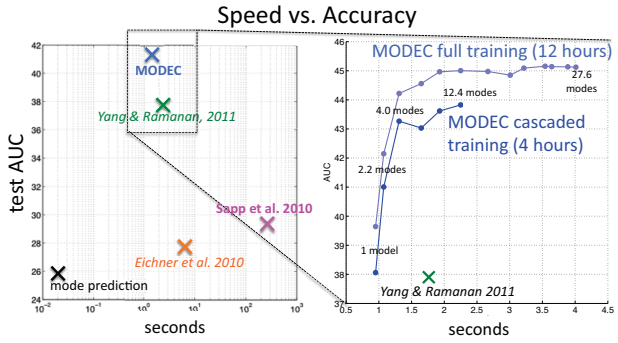

Figure 5. Test time speed versus accuracy. Accuracy is measured as area under the pixel error threshold curve (AUC), evaluated on the FLIC testset. Speed is in seconds on an AMD Opteron 4284 CPU @ 3.00 GHz with 16 cores. See text for details.

Qualitative results. Example output of our system on the FLIC test set is shown in Figure 6.

8. Conclusion

We have presented the MODEC model, which provides a way to maintain the efficiency of simple, tractable models while gaining the rich modeling power of many global modes. This allows us to perform joint training and inference to manage the competition between modes in a principled way. The results are compelling: we dominate across the accuracy-speed curve on public and new datasets, and demonstrate the importance of multimodality and efficient models that exploit it.

Acknowledgments

The authors were partially supported by ONR MURI N000141010934, NSF CAREER 1054215 and by STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. CVPR*, 2009.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Proc. ICCV*, 2009.

[3] K. Duan, D. Batra, and D. Crandall. A multi-layer composite model for human pose estimation. In *Proc. BMVC*, 2012.

[4] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *Proc. BMVC*, 2009.

[5] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Articulated human pose estimation and search in (almost) unconstrained still images. Technical report, ETH Zurich, D-ITET, BIWI, 2010.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results, 2009.

[7] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. A library for large linear classification. *JMLR*, 2008.

Figure 6. Qualitative results. Shown are input test images from FLIC. The top 4 modes unpruned by the cascade step are shown in order of their mode scores z_ℓ, z_r on the left and right side of each image. The mode chosen by MODEC is highlighted in green. The best parse $y^\star$ is overlaid on the image for the right (blue) and left (green) sides. In the last row we show common failures: firing on foreground clutter, background clutter, and wrong scale estimation.

[8] Felzenszwalb, Girshick, and McAllester. Discriminatively trained deformable part models, release 4, 2011.

[9] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[10] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *Proc. ICCV*, 2007.

[11] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proc. CVPR*, 2011.

[12] L. Ladicky and P. H. Torr. Locally linear support vector machines. In *ICML*, 2011.

[13] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *Proc. ICCV*, 2011.

[14] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *Proc. CVPR*, 2006.

[15] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *Proc. ECCV*, 2010.

[16] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *Proc. CVPR*, 2011.

[17] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *Proc. ICCV*, 2011.

[18] Y. Tian, C. Zitnick, and S. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *Proc. ECCV*, 2012.

[19] D. Tran and D. Forsyth. Improved Human Parsing with a Full Relational Model. In *Proc. ECCV*, 2010.

[20] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proc. ECCV*, 2008.

[21] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *Proc. CVPR*, 2011.

[22] D. Weiss, B. Sapp, and B. Taskar. Structured prediction cascades (under review). In *JMLR*, 2012.

[23] D. Weiss and B. Taskar. Structured prediction cascades. In *Proc. AISTATS*, 2010.

[24] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *Proc. CVPR*, 2011.

[25] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *Proc. CVPR*, 2012.