Human Pose Estimation: Extension and Application

Digvijay Singh

CVIT, IIIT Hyderabad

Guide: Professor C V Jawahar

Content

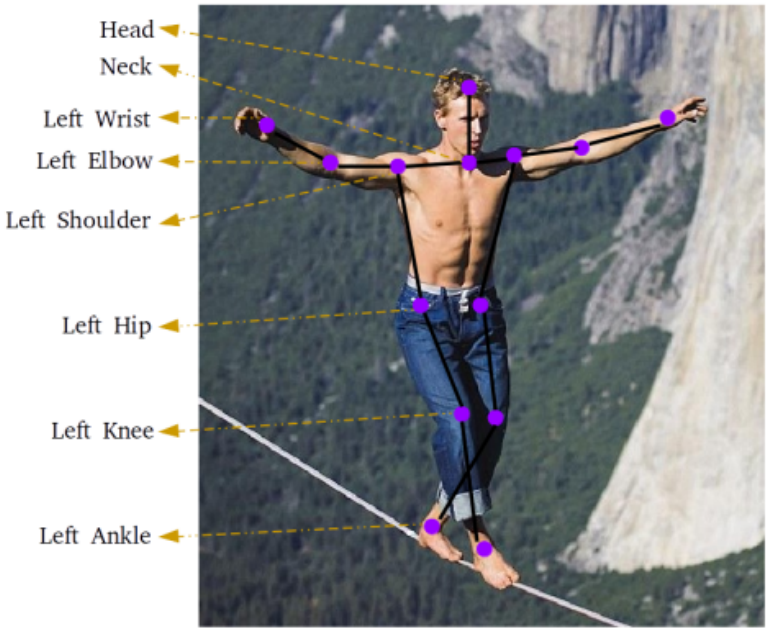
1. Problem Definition and Challenges
2. Previous work on Human Pose Estimation: Classic to New
3. **Extension**: Human Pose Estimation in Videos
4. **Application**: Parsing Clothes in Images

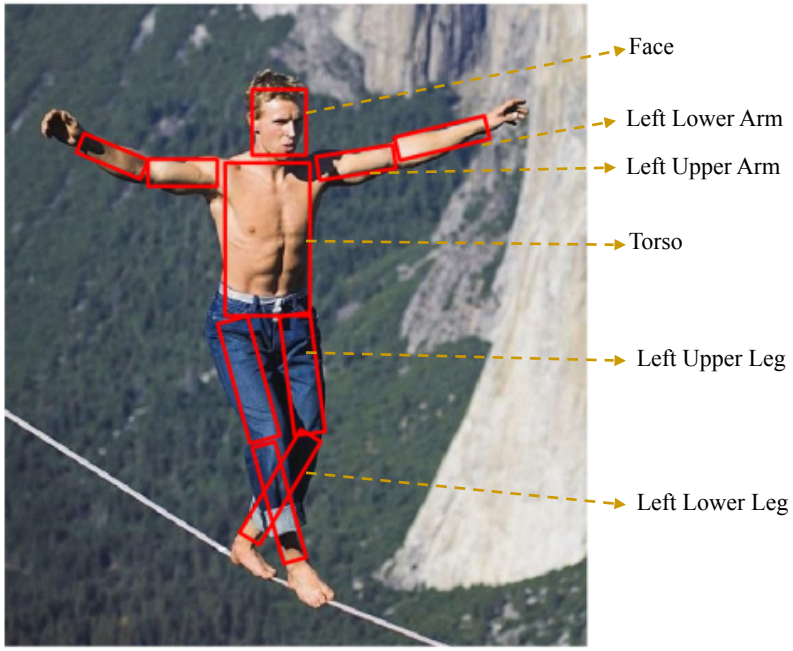
HPE: Problem Definition

Locating Body Joints

OR

Locating Body Parts





Motivation

Most studied subject : HUMAN

Where is the human located?
How many humans?
What activity is the human performing?
What gesture humans are making?

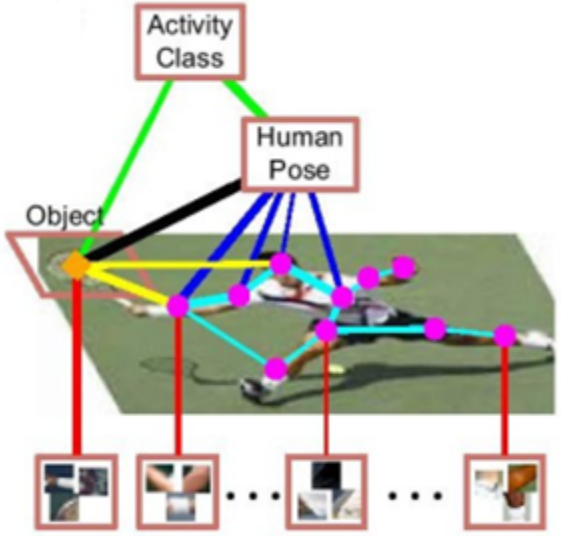
Activity: People toasting.
8 humans: 4 male, 4 female.
Background: Club
Interaction: Holding glasses

Activity: Boys playing football.
8 humans: 8 boys
Background: Field
Interaction: Kicking football

Motivation for HPE

Detecting Cloth Segments [1]

Understanding Human-Object [2] interaction and performed Activity



1 Yamaguchi et al, Parsing Clothing in fashion photographs.
2 Yao et al, Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activites.

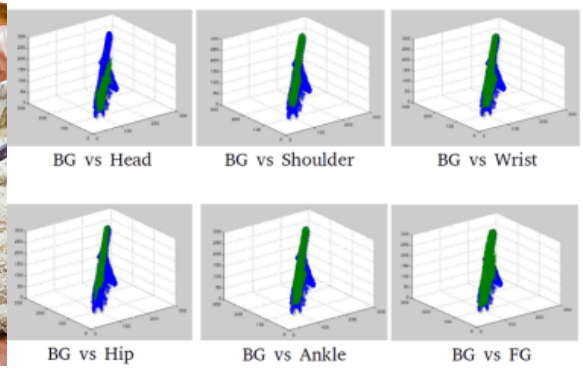
Challenges for HPE

Clothing Variations

Different clothing induces different silhouette.
Different clothing covers/reveals different body regions.

Illumination Variation

Outdoor illumination is highly variable.



Background Clutter

Background comprises of wide range of color distribution.

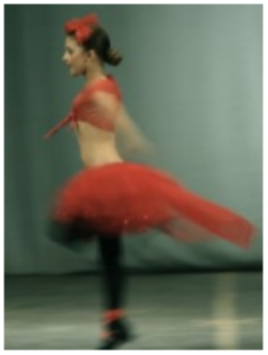
Challenges for HPE

Occlusion: Self and External

Natural settings can have occlusions induced by different entities present.

Body part Foreshortening

Body parts projecting with a higher angle with respect to image plane have foreshortened projection on image plane.



Motion Blur (Video data specific)

Quick moving body parts cause blur while capturing.

Our Contribution

Extending Human Pose Estimation for Videos

- Uses off-the-shelf pose estimator for images.
- Learns video-specific features using semi-supervised algorithm.
- Iterative self-training propagates correctness across sequence which is more robust than using contemporary tracking strategies.

Applying HPE for Cloth Parsing in Images

- By incorporating more robust pose information, the model is shown to perform more reliably in unrestricted settings.
- A peek at relevant information extraction from obtained results like obtaining cloth type/color patterns usually worn.

Content

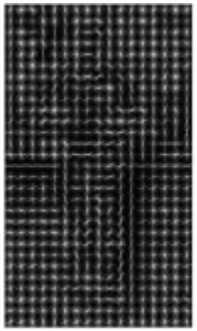
1. Problem Definition and Challenges
2. Previous work on Human Pose Estimation: Classic to New
3. **Extension**: Human Pose Estimation in Videos
4. **Application**: Parsing Clothes in Images

Previous work on HPE: Classic

1. Histogram of Gradients (HoG)

Image divided into dense grid of uniformly spaced cells.

Gradient orientations are accumulated in all the grid cells.



HoG and Human Detection

Simple linear SVM trained using HoG features gives reliable pedestrian detector.

* Dalal et al. Histograms of Oriented Gradients for Human Detection.

Previous work on HPE: Classic

2. Poselets: Body Part Detectors trained with Clustering in configuration space

Keypoint based similarity distance: $D(a,b) = w_a . ||x_a - x_b||_2^2 . (1 + v(a,b))$

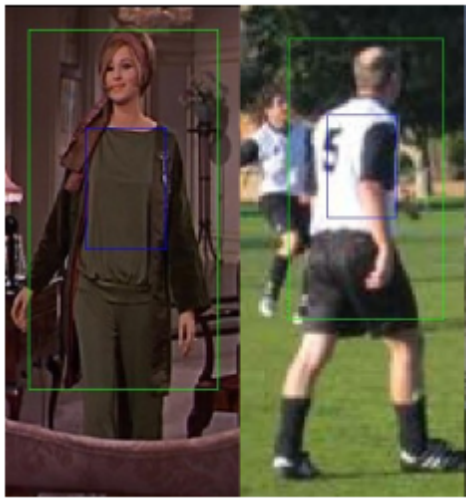
x_a, x_b: 3D coordinates of sample a, b ;
$v(a,b)$: Penalizes visual dissimilarity

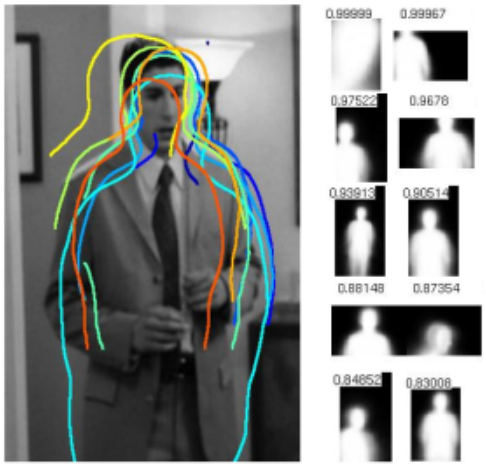
* Bourdev et al, Poselets: Body Part Detectors trained with Clustering in configuration space.

Previous work on HPE: Classic

Poselets (contd.)

Training: Poselet classifiers are linear SVM trained using poselet examples as positives and non-person patches as negatives.
Features used are HoG.





Inference: All poselet classifiers used to get probability maps for different body parts. Probability maps are combined using optimal max-margin framework to get detection box.

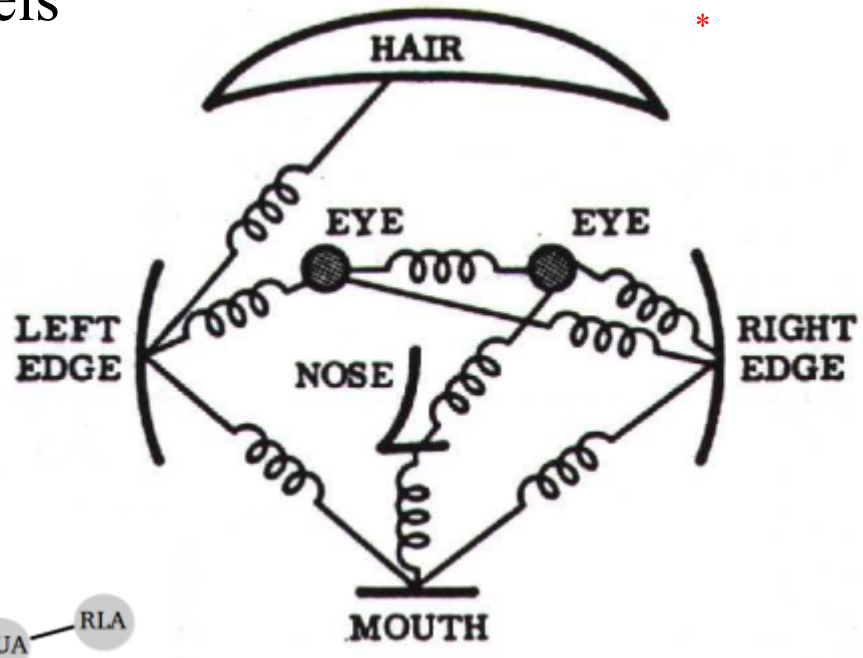
Pros: Each part classifier independent of the presence/absence of other parts.
Cons: Fails to exploit holistic representation of human pose.

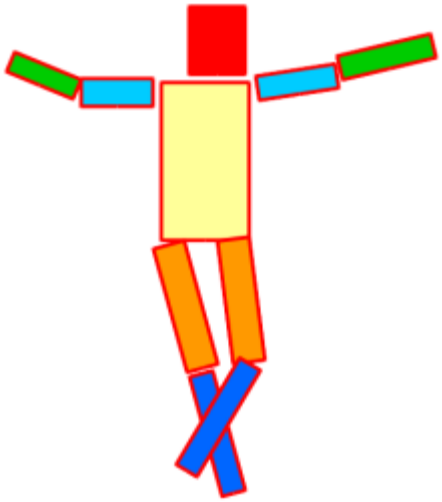
Previous work on HPE: Classic

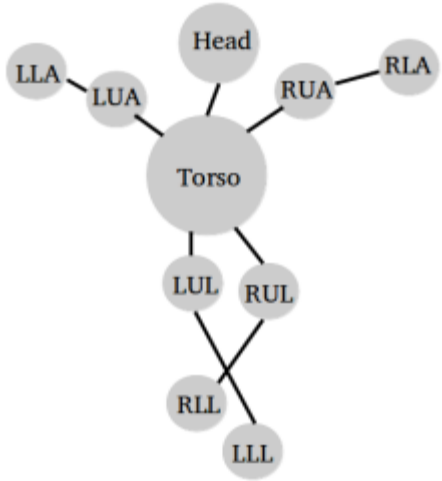
3. Pictorial Structure based Models

For entities that are an embedding of different sub-parts.

Example: Face and its sub-parts: eyes, nose, ears, mouth.



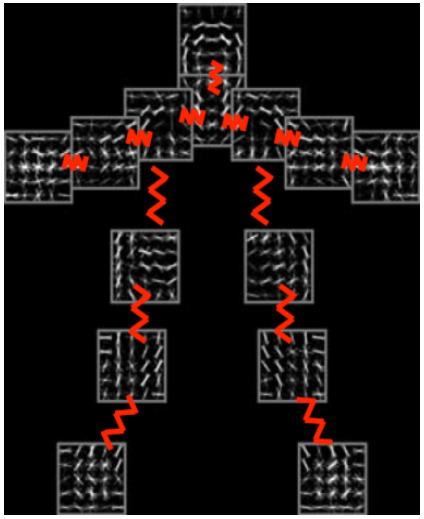




* Fischler et al, The Representation and Matching of Pictorial Structures.

Previous work on HPE: Classic

Pictorial Structure for Human Pose : YR Model



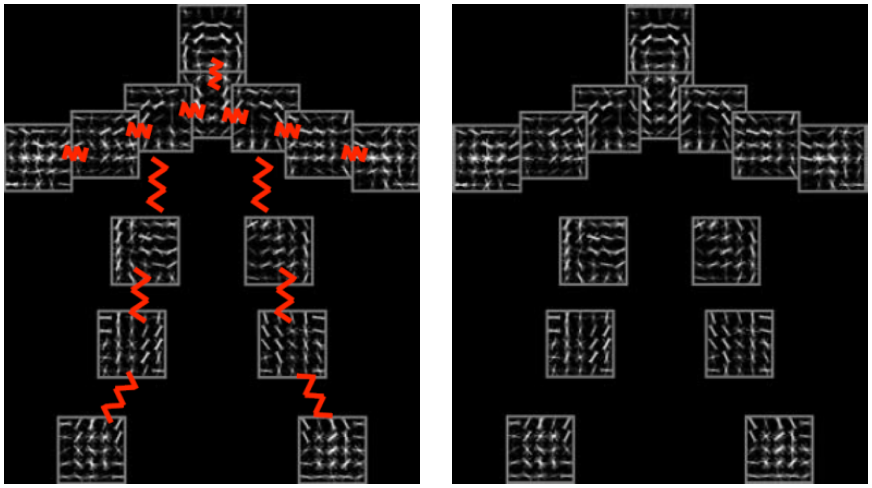
$$S(I, L)$$

- I: Image
- L_i: Location of part i

Slide Credits: Yang et al, Articulated Human Detection with Flexible Mixtures-of-Parts.

Previous work on HPE: Classic

Pictorial Structure for Human Pose : YR Model



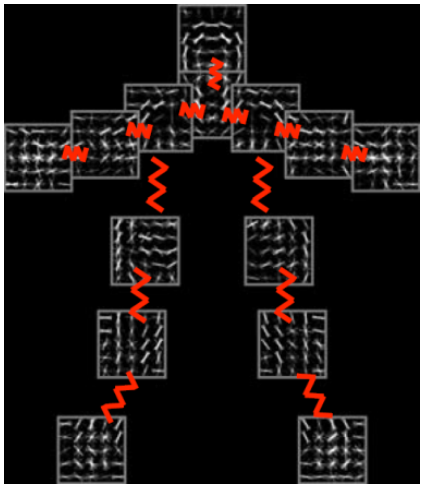
$$S(I, L) = \sum_{i \in V} \alpha_i \cdot \phi(I, l_i)$$

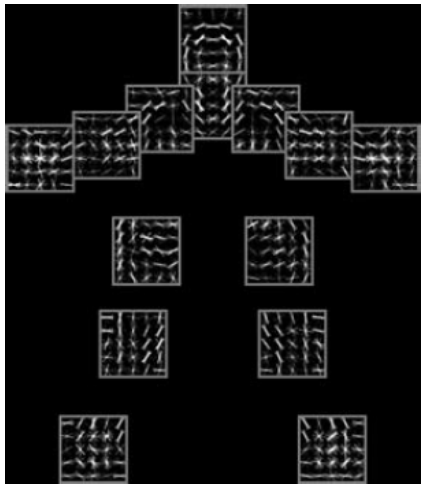
- α_i: Unary template for part i
- $\phi(I, l_i)$: Local image features at location i

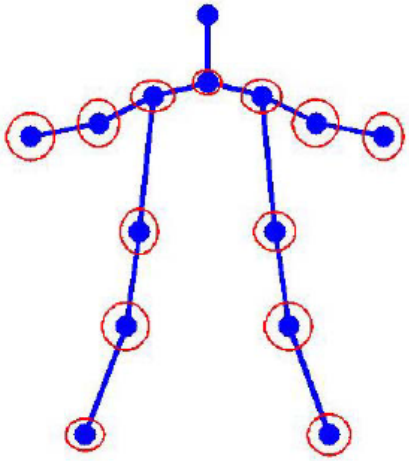
Slide Credits: Yang et al, Articulated Human Detection with Flexible Mixtures-of-Parts.

Previous work on HPE: Classic

Pictorial Structure for Human Pose : YR Model







$$S(I,L) = \sum_{i \in V} \alpha_i \cdot \phi(I, l_i) + \sum_{ij \in E} \beta_{ij} \cdot \psi(l_i, l_j)$$

- $\psi(l_i, l_j)$: Spatial features between l_i and l_j
- β_{ij} :Pairwise springs between parts i and j

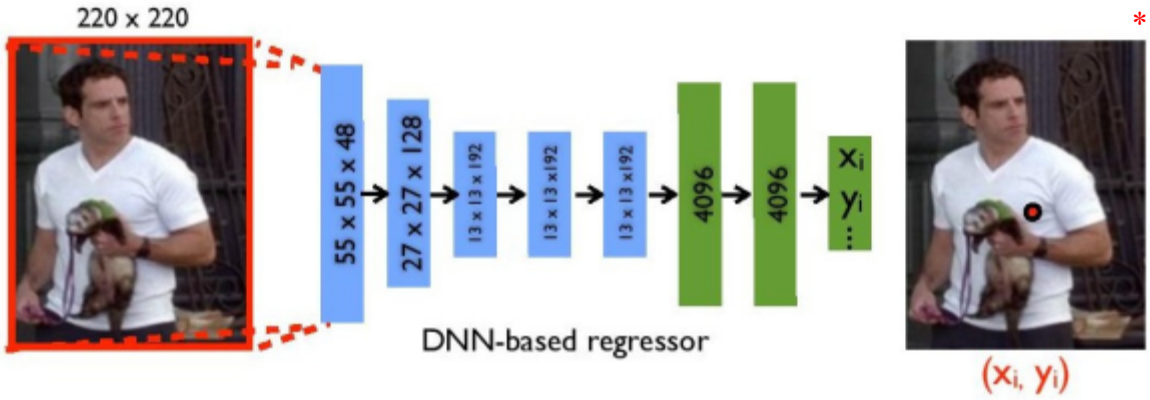
Slide Credits: Yang et al, Articulated Human Detection with Flexible Mixtures-of-Parts.

Previous work on HPE: New

4. DeepPose: Human Pose Estimation via Deep Neural Networks

For k keypoints, the architecture is designed to regress a vector of size $2k$

$$LossValue(L_2) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{k} ||y_i - \varphi(x_i, \theta)||_2^2$$



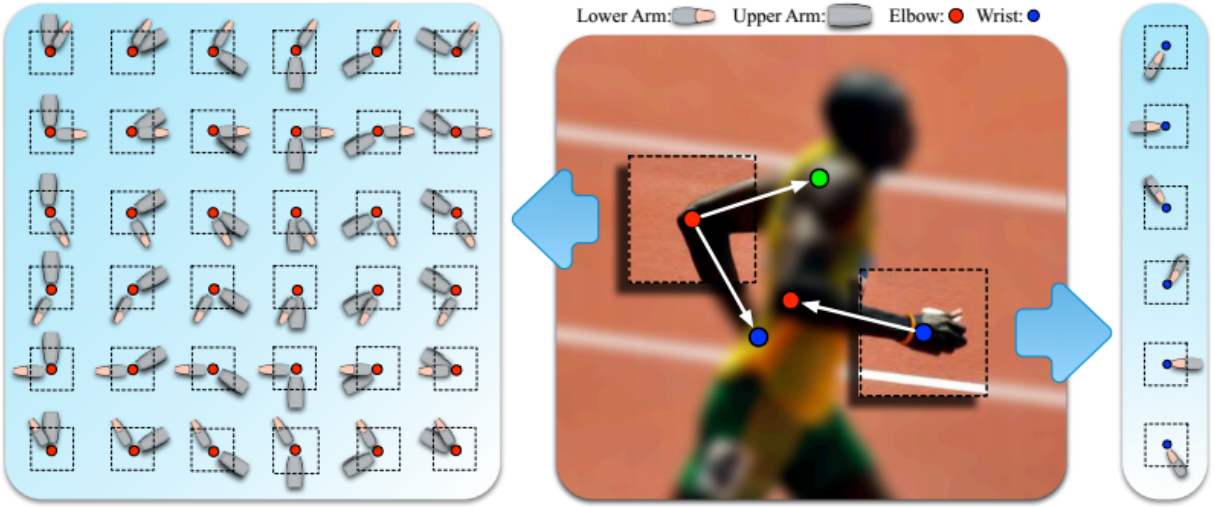
Parameters φ are fine-tuned using stochastic gradient descent while back-propagation.

* Toshev et al, DeepPose: Human Pose Estimation via Deep Neural Networks.

Previous work on HPE: New

5. Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations (IDPR)

Assumption: Local image features around a part can reliably detect the part as well as help in determining relative positions of all its neighbors.



*

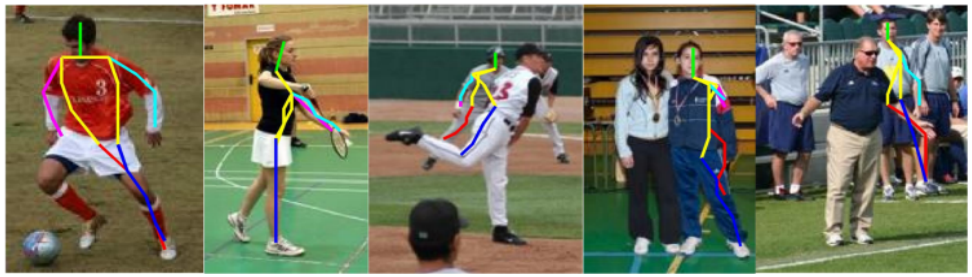
* Chen et al, Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations.

Classic vs New: Comparison

FLIC dataset

Method	PCK(%)	Examples
YR	70.00	
IDPR	91.14	

LSP dataset

Method	PCK(%)	Examples
YR	55.81	
IDPR	67.91	

Content

1. Problem Definition and Challenges
2. Previous work on Human Pose Estimation: Classic to New
3. **Extension**: Human Pose Estimation in Videos
4. **Application**: Parsing Clothes in Images

Drawbacks of existing work

- Pairwise over-rides unary response to make false predictions.

Drawbacks of existing work

- Pairwise over-rides unary response to make false predictions.
- Very high number of possible configurations for body parts that needs higher parameterization.
- Contemporary tracking methods are restricted by the amount of object movement and outlook change it can handle.

Problem Formulation

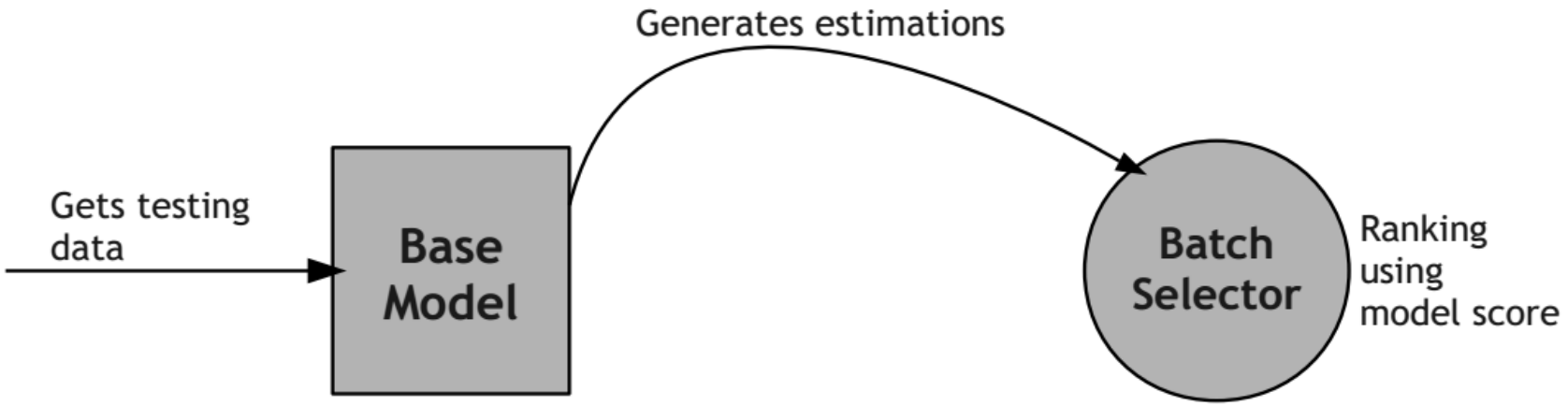
- Video data has consistent appearance and slow changing pose configuration.

Problem Formulation

- Video data has consistent appearance and slow changing pose configuration.
- While temporal consistency can be learnt from supervised data, modelling unseen video-specific appearance is difficult.
- We propose to augment off-the-shelf HPE models with additional parameters to encode local appearance.
- These additional parameters are learnt using confident pose estimations for a particular video.
- Final model is run on all the frames in the video.

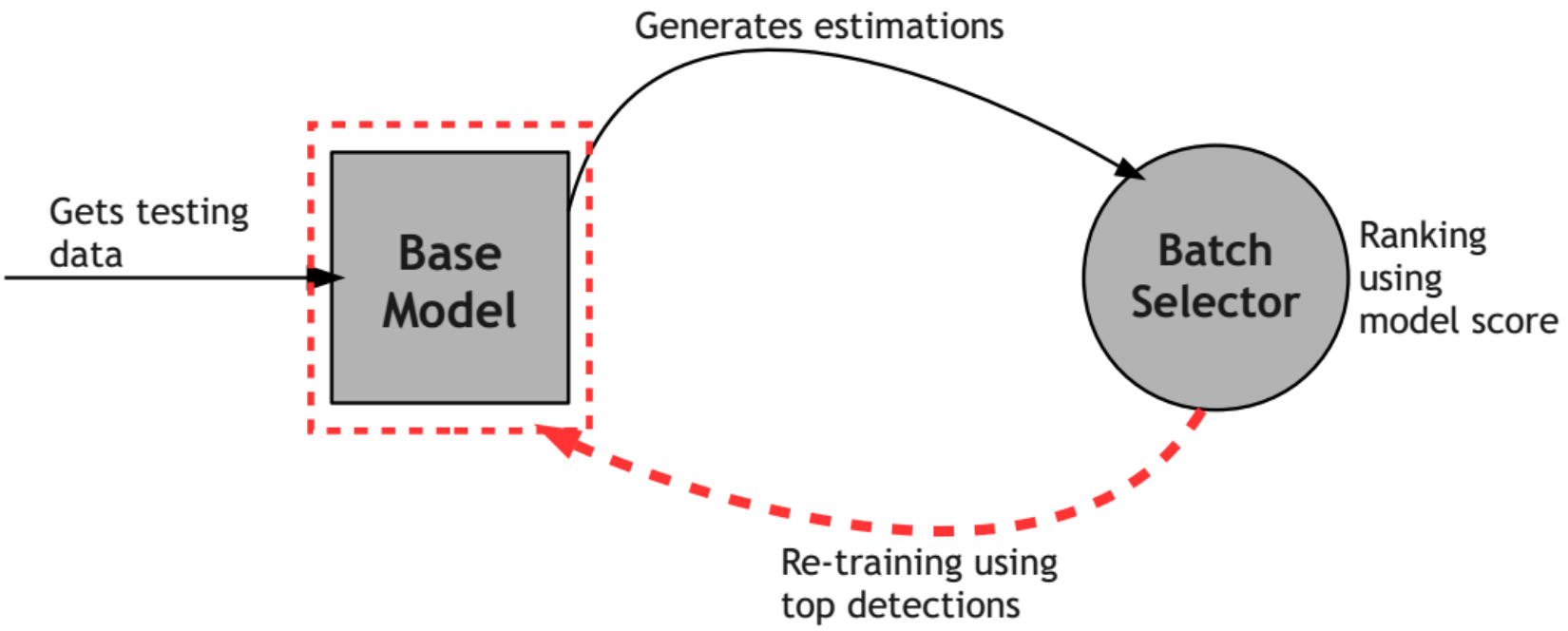
Proposed Method

Standard Self-training Pipeline:



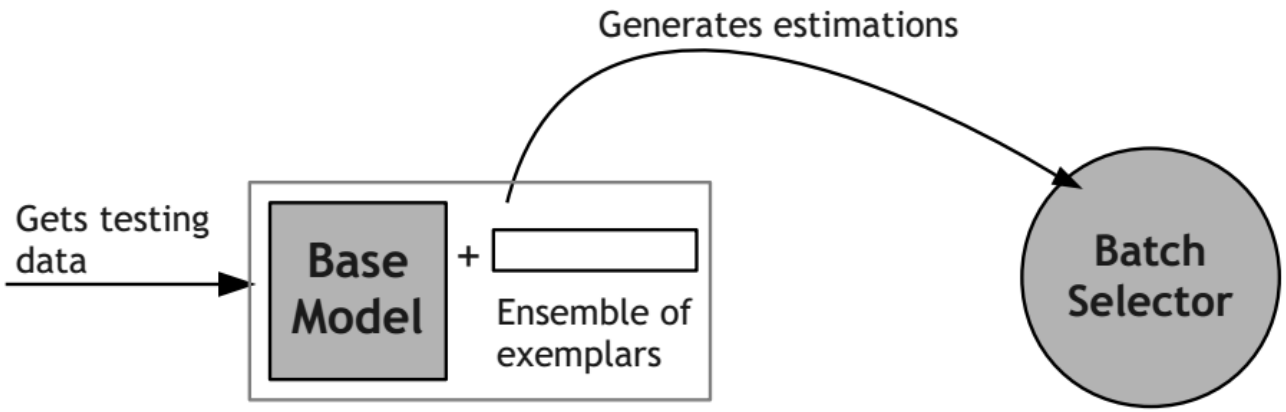
Proposed Method

Standard Self-training Pipeline:



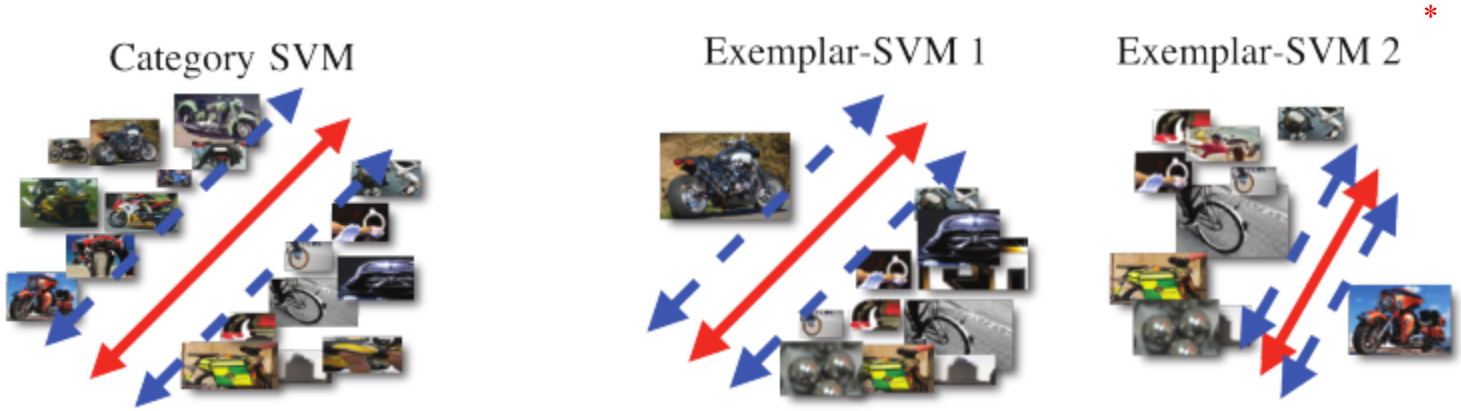
Proposed Method

Modified Self-training Pipeline:



Exemplar-SVM

While training, each exemplar generates a classifier with exemplar as only positive and large amount of negatives not belonging to exemplar's class.



*

$$\Omega_E(\mathbf{w}, b) = ||\mathbf{w}||^2 + C_1 h(\mathbf{w}^T \mathbf{x}_E + b) + C_2 \sum_{\mathbf{x} \in \mathcal{N}_E} h(-\mathbf{w}^T \mathbf{x} - b)$$

h(x) = max(1-x,0) "hinge-loss"

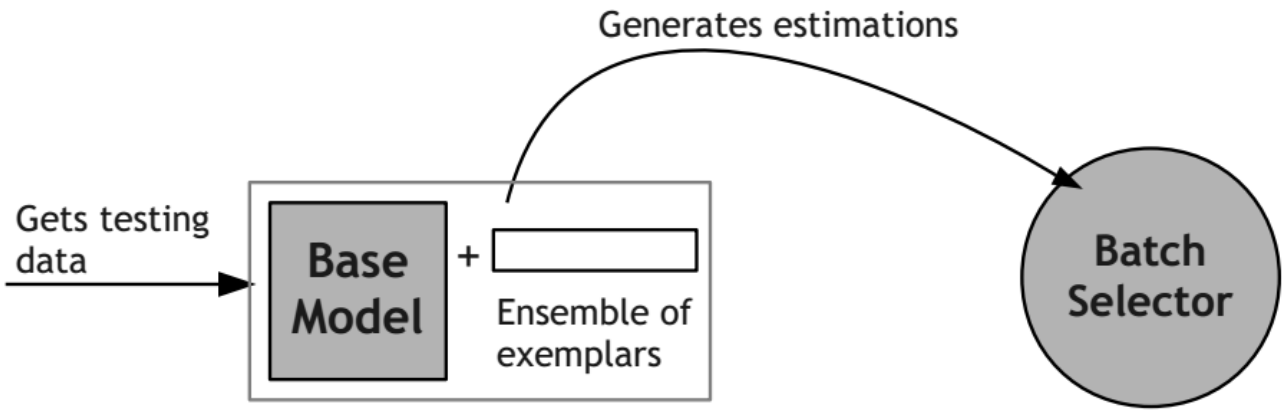
The method exploits:

- Effectiveness of discriminative detectors
- Explicit nearest-neighbor correspondence for instances from similar class.

* Malisiewicz et al, Ensemble of Exemplar-SVMs for Object Detection and Beyond.

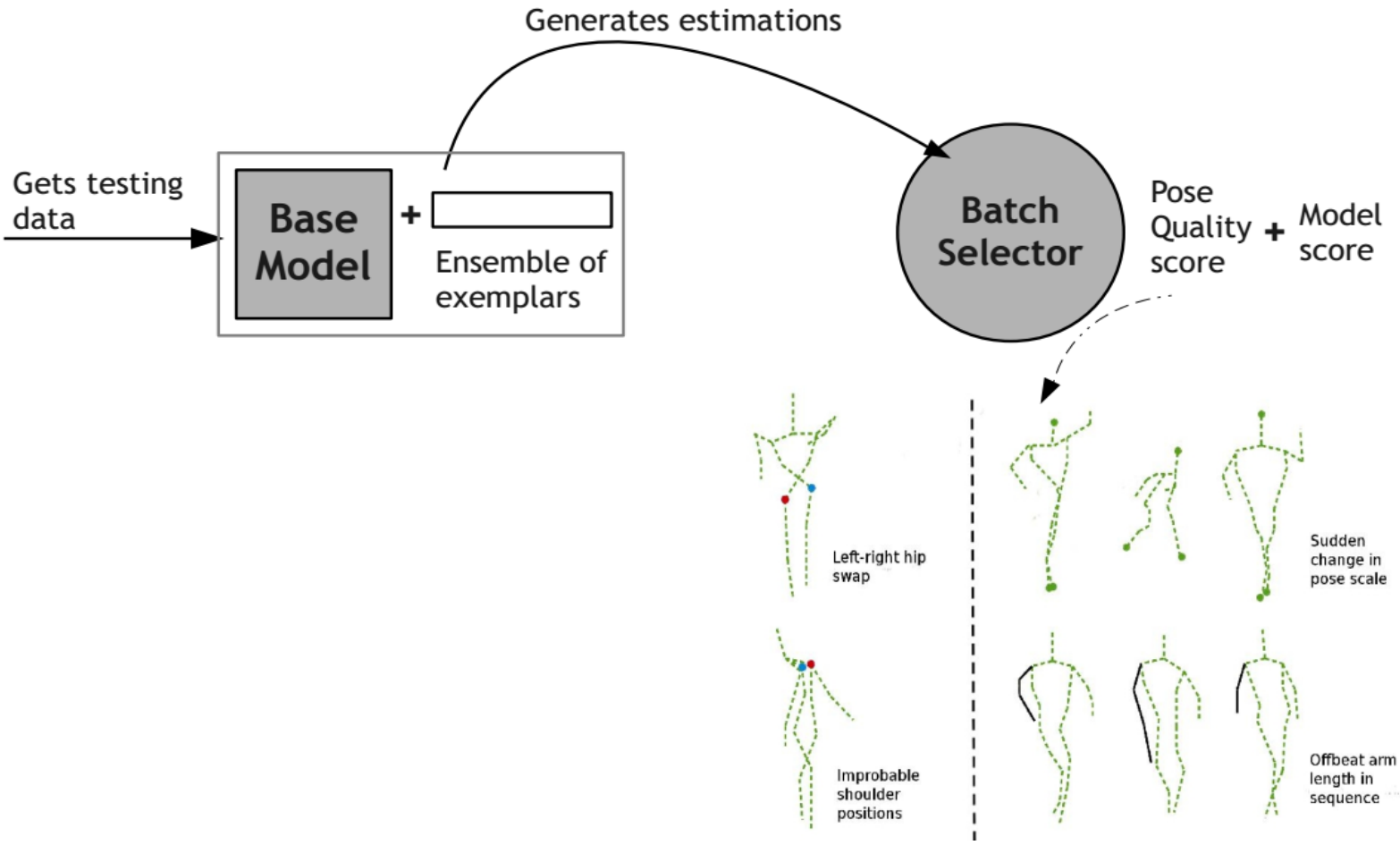
Proposed Method

Modified Self-training Pipeline:

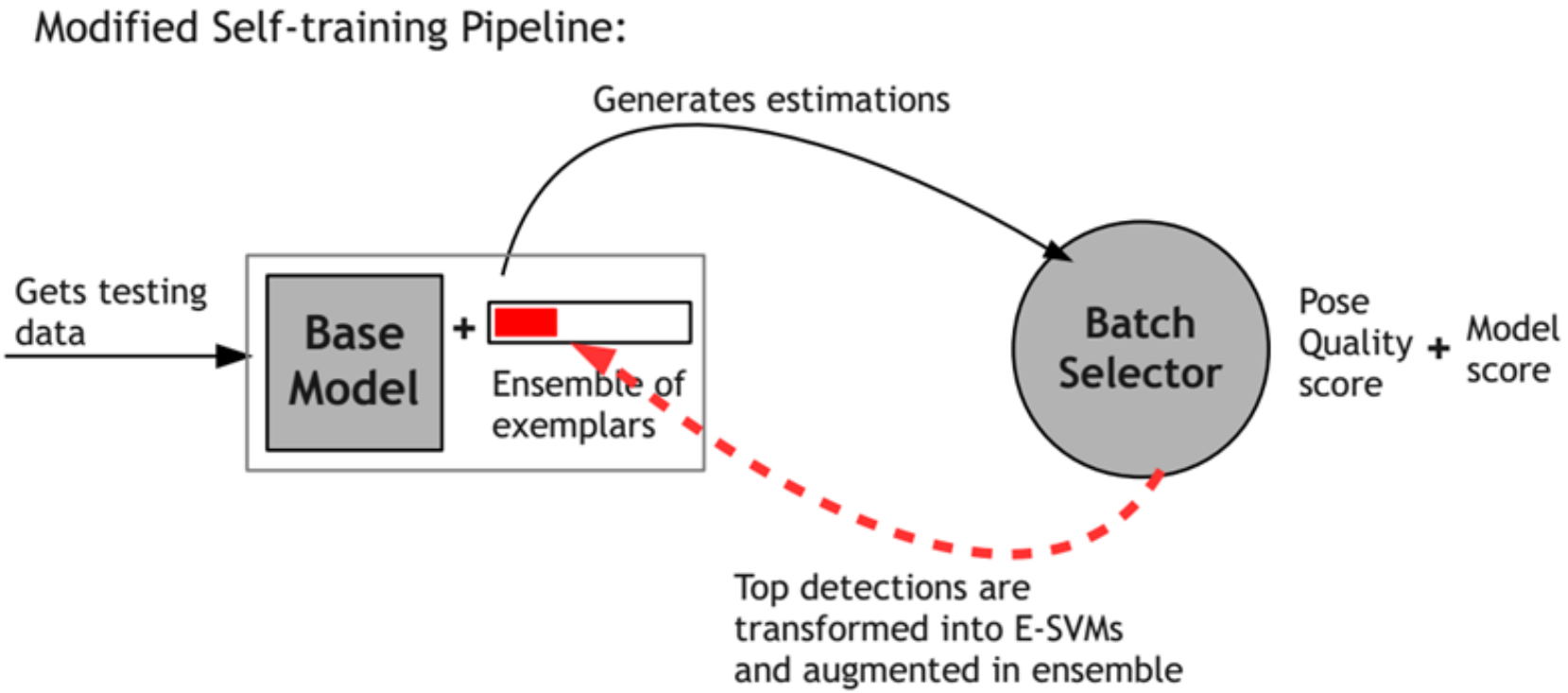


Proposed Method

Modified Self-training Pipeline:

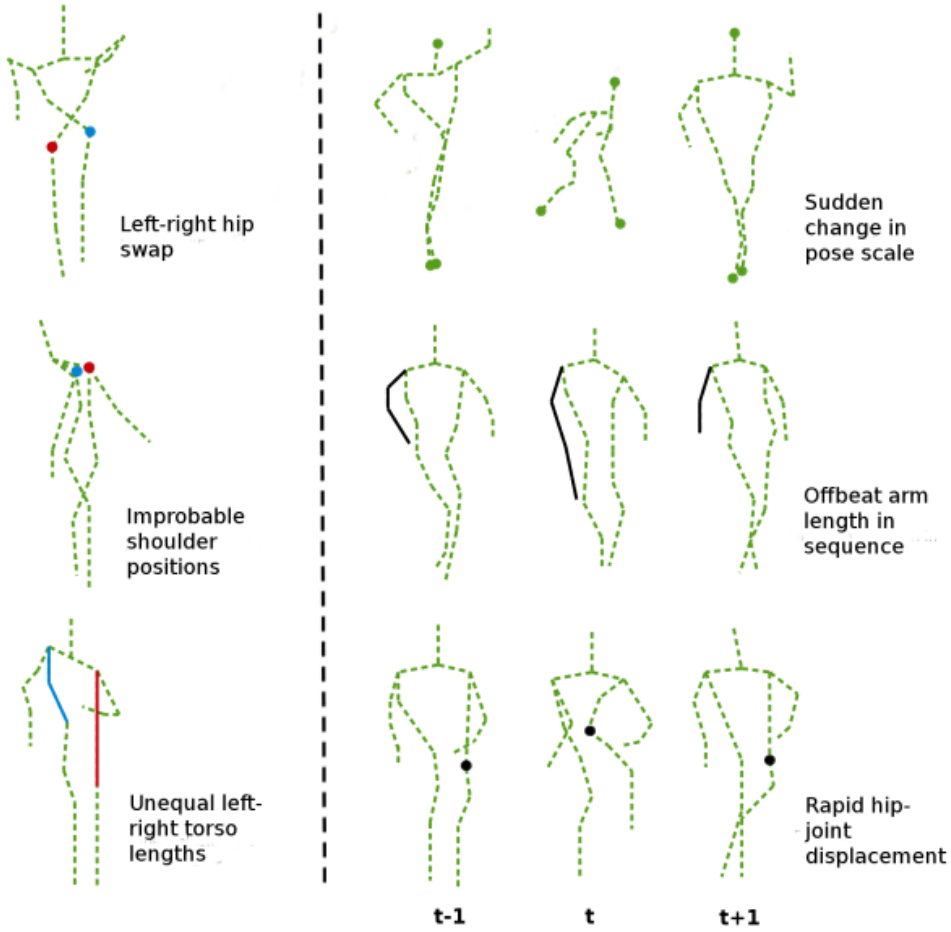


Proposed Method



Pose Quality Ranking SVM

Based on the characteristics of feasible pose geometric configurations in current frame as well as local sequence of frames.



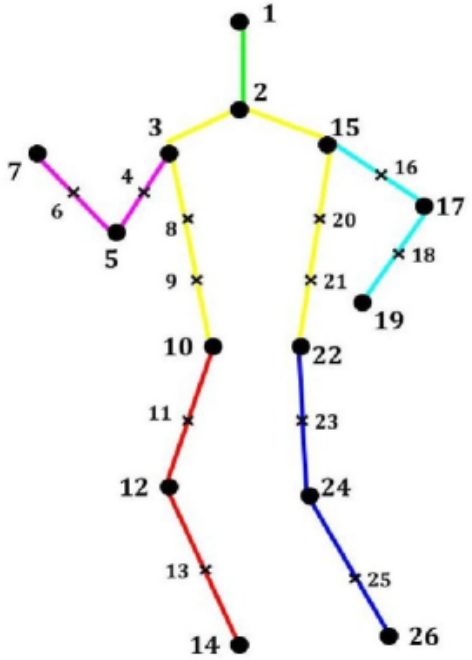
Pose Quality Ranking SVM

Per frame criteria:

- Left-Right shoulder $x(3) < x(15)$
- Left-Right hip $x(10) < x(22)$
- Left-Right torso length $t1 < d(3,10)/d(15,22) < t2$

Temporal neighborhood consistency criteria:

- Pose scale $t1 < sc(i)/nbd(sc(i)) < t2$
- Left shoulder movement $t1 < d(x(3),nbd(x(3)) < t2$
- Right hip movement $t1 < d(x(22),nbd(x(22)) < t2$

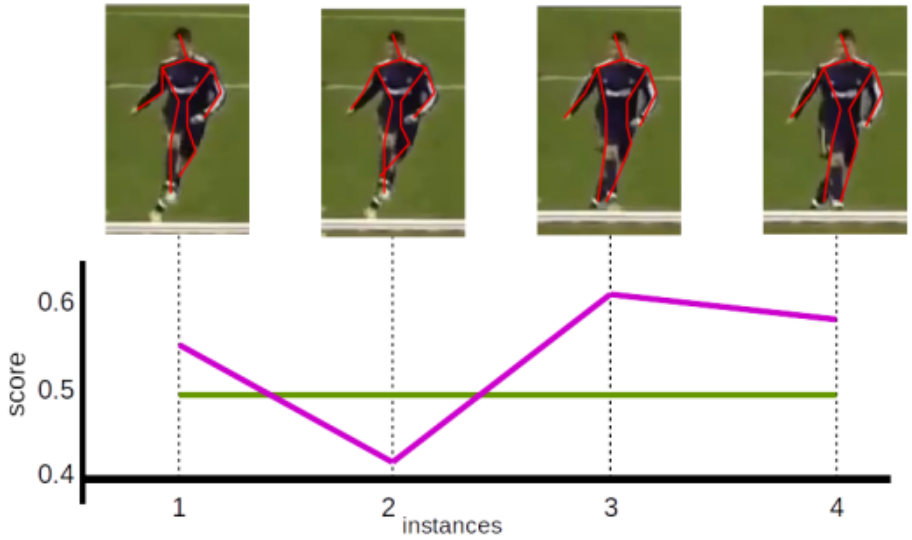


Coarse-to-Fine Strategy for Exemplar Selection

To ensure samples picked in each iteration of self-training belong to different temporal segments.

Strategy:

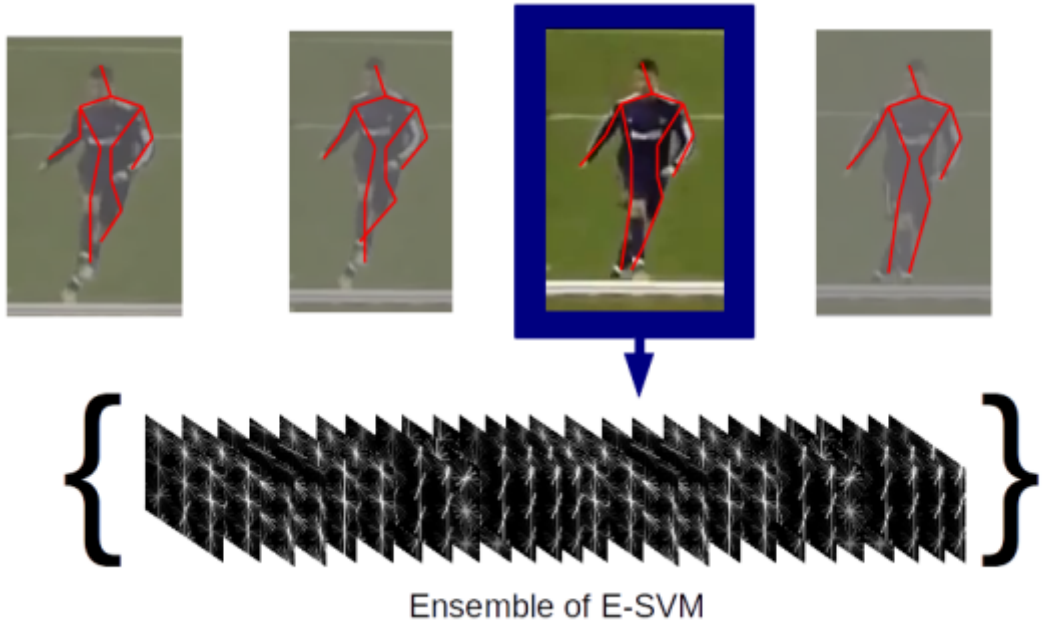
- Temporal regions with mean pose quality score higher than mean pose quality score of full sequence are chosen.
  Coarsely, *k* regions with highest mean output scores are picked.
- From each selected temporal region, maximum *j* exemplars with highest pose quality score are finally picked.



Ensemble of Exemplar SVMs for Semi-Supervised Self-Training

Exemplars picked using our strategy are converted into Exemplar-SVMs.

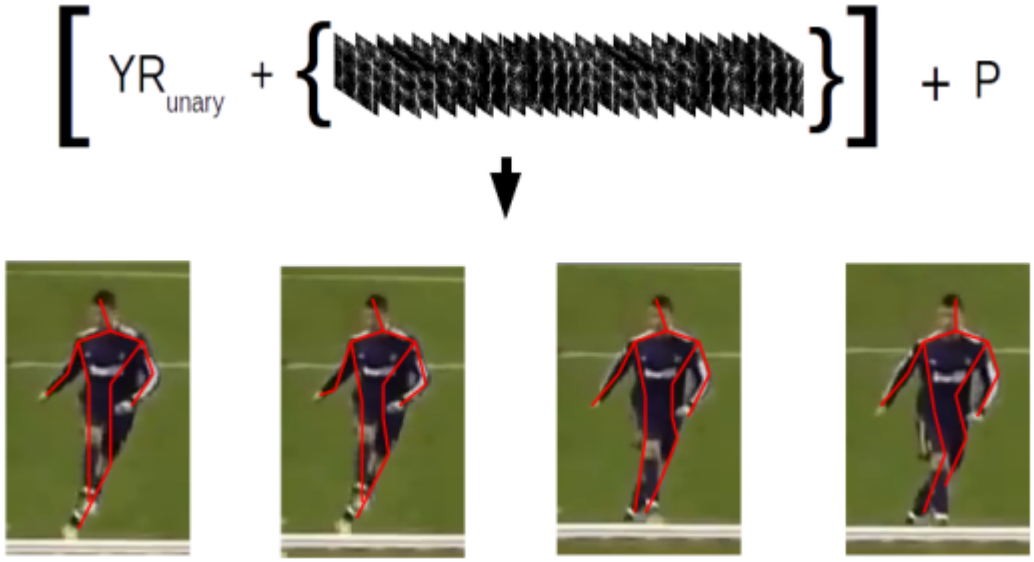
For full-pose, 26 parts synthesize 26 E-SVMs, with body-part as the only positive and random patches from non-human image as negatives.



Ensemble of Exemplar SVMs for Semi-Supervised Self-Training

Unary potential term is redefined as: $\sum_{i \in V} (\eta w_i^{t_i} + (1-\eta)\hat{w}_i^{t_i}) \cdot \phi(I, p_i)$

where $\hat{w}_i^{t_i}$ are the (normalized) weights learnt from the exemplar SVM for the i^{th} part tuned for type t_i, and η is a parameter that controls the weight given to the exemplars' contribution to the unary score.



Evaluation Metric and Datasets used

Percentage of Correct Keypoints (PCK):
Estimated keypoints are correct that lie within a threshold distance D_t from its GT counterpart.

$$D_t = \beta . \max(h, w)$$

Datasets Used

Dataset	Avg Frame length	Activities	Complexity
VideoPose	30	TV shows	Easy
Poses in the Wild	30	Movies	Medium
CVIT-Sports-Videos	131	Sports	High

Datasets Used

CVIT-Sports-Videos

Poses in the Wild

VideoPose

CVIT-Sports-Videos

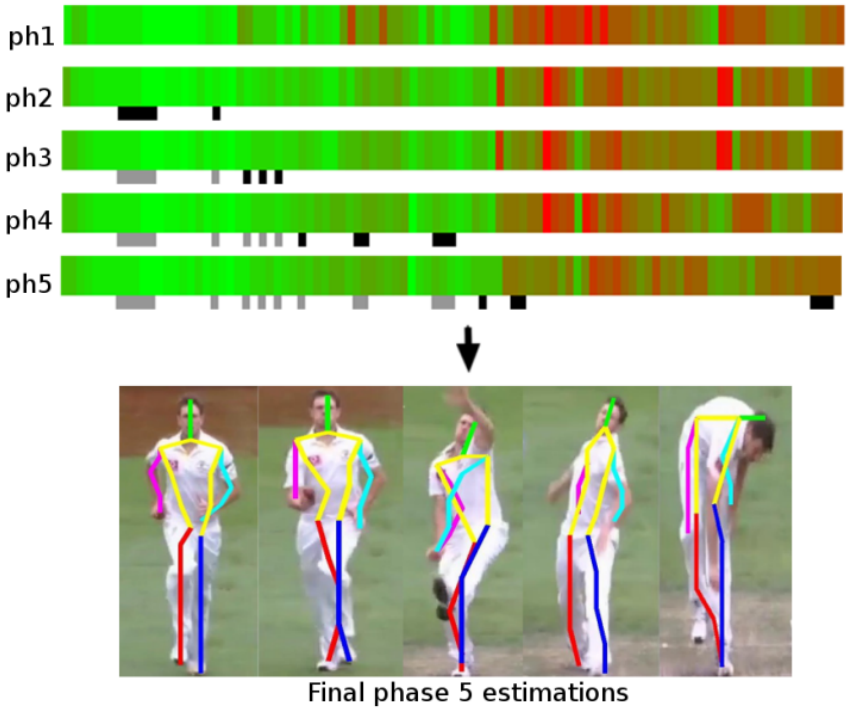
- 11 sports videos retrieved from YouTube.
- Activities included: cricket-bowling, cricket-batting and football-kicking.
- Full human pose (26 keypoints) labelling
- Total 1446 frames averaging to 131 frames per video.

Half-body self occlusion

Extreme body deformation

Quantitative Results

Method	PCK(%)
YR	72.41
FT @ phase2	74.31
FT-Full	74.19



Final phase 5 estimations

Quantitative Results

Method	PCK(%)
YR	72.41
FT @ phase2	74.31
FT-Full	74.19
FT-Full + PP	**74.74**

PP, Post Processing: Neighborhood Interpolation

$$est_i = mean(est_{i-1}, est_{i+1})$$

Quantitative Results

Method	PCK(%)
YR	72.41
FT @ phase2	74.31
FT-Full	74.19
FT-Full + PP	**74.74**

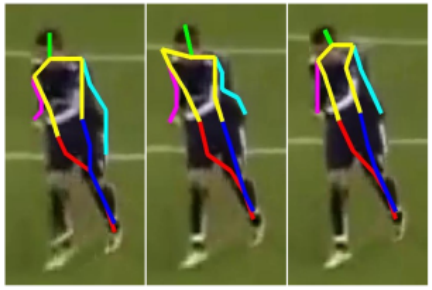
Full-body evaluation

Upper-body evaluation

Method	CVIT-Sports-Videos	PIW	VP
YR	64.87	70.74	63.35
[1]	42.12	71.43	**71.95**
YR + SiftFlow	31.75	48.04	60.01
FT @ phase 2	**65.20**	72.78	69.11
FT-Full	64.90	72.44	68.65
FT-Full + PP	64.50	**73.26**	68.96

[1] Cherian et al, Mixing body-part sequences for Human Pose Estimation.

Qualitative Results



CVIT-Sports-Videos

Top Row: YR
Bottom Row: FT-Full

Poses in the Wild

Discussions

- Proposed self-training methodology that empowers unary response and captures more intricate pose configurations.
- Introduced new CVIT-Sports-Videos dataset having 11 videos from sports domain.
- Novel pose quality scoring criteria that helps in selecting instances for each iteration of self-training.
- Quantitative and qualitative results show we surpass previous state-of-the-art based on pictorial structures.
- **Parameter Selection:** The value of biasing parameter η that controls the contribution from base model and exemplars is determined automatically using pose quality score in a manner determined by cross validation.

Content

1. Problem Definition and Challenges
2. Previous work on Human Pose Estimation: Classic to New
3. **Extension**: Human Pose Estimation in Videos
4. **Application**: Parsing Clothes in Images

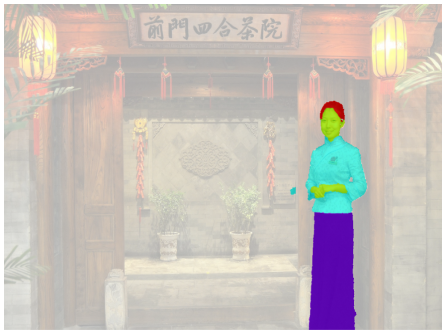
Application: Clothing Parsing

Simple Image

Complex Pose

Complex Occlusion

Complex Clothing Patterns



Lower Clothes
Upper Clothes
Skin
Hair
Null

Lower Clothes
Upper Clothes
Skin
Hair
Null
Occluding Object
Shoes

Skin
Lower Clothes
Upper Clothes
Hair
Null
Occluding Object

Hat
Upper Clothes
Skin
Dress
Null
Occluding Object

Application: Clothing Parsing

Understanding the dependence of obtaining cloth segments and determining cloth types on **pose information**. Settings are divided into two categories:

Restricted

- Front-facing single human present in the center of the image with similar scale and width.
- Pose configurations are very generic and comprehensible.
- Almost all human body parts are clearly visible
- Dataset: Fashionista

Unrestricted

- No restrictions on orientation, scale or number of humans present in the image.
- Pose configuration composition can vary from very simple to very intricate.
- No restriction on missing human body parts.
- Dataset: H3D

Application: Clothing Parsing

Understanding the dependence of obtaining cloth segments and determining cloth types on **pose information**. Settings are divided into two categories:

Restricted

More assumptions.

Feasibility of a more regularized model.

Pictorial Structure based models like YR model work reliably well under such assumptions.

Tackled by Yamaguchi et al.

Unrestricted

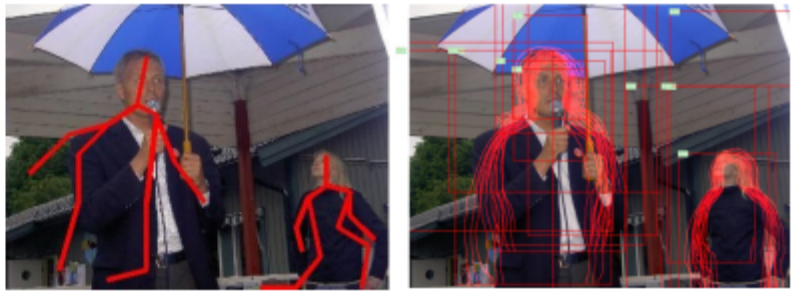
Fewer assumptions.

More random variables to keep account of.

Poselets (body part detector) works better in such settings.

Our problem.

Robust Pose Information



Pictorial Structure based models underperform in natural settings.
Poselets manage to approximate visible body-parts.

Given prior information, certain assumptions can be made

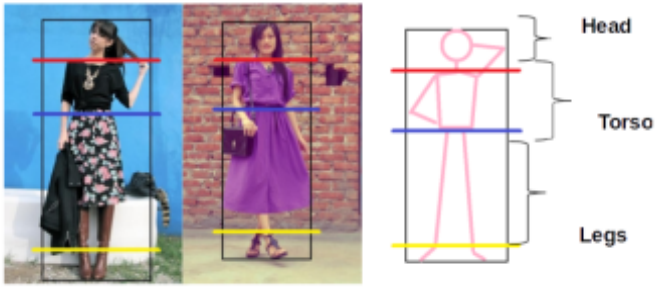
Cloth & body part correlation

Sweater-Torso Pants-Legs

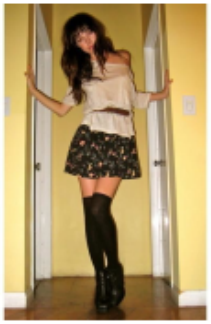
Cloth & appearance correlation

Skirt Skin

Cloth & position correlation



Application: Parsing Clothes in Unrestricted Images



Superpixels

Features

(1) RGB Histogram (2) CIE L*a*b Histogram (3) Gabor filter response (4) Normalized x-y coordinates histogram (5) Normalized x-y coordinates w.r.t body joints.

CRF

$$L^* \equiv \min_L \sum_{i \in V} \Phi(l_i \mid Z, \phi(s_j, Z)) + \lambda_1 \sum_{(i,j) \in E} \psi_1(l_i, l_j) + \lambda_2 \sum_{(i,j) \in E} \psi_2(l_i, l_j \mid \psi(s_i, s_j, Z))$$

where **Z** is the human pose information, **I** is the image and

$$\psi(s_i, s_j, Z) \equiv [(\phi(s_i, Z) + \phi(s_j, Z))/2, |(\phi(s_i, Z) - \phi(s_j, Z))/2|]$$

Segmentation

ϕ is the feature vector

Φ is the unary potential modeling the appearance

Ψ_1 is the pairwise potential modeling the label co-occurrence

Ψ_2 is the pairwise potential modeling the appearance dependent co-occurrence

Learning: Logistic regression

Inference: Loopy belief propagation

Datasets Used

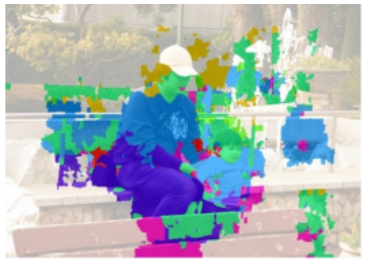
H3D

Fashionista

Qualitative Results

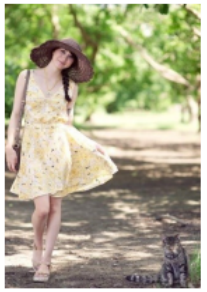







H3D





Fashionista

Quantitative Results

Method	Full-a		Full-m		Unary	
	Pixel acc	mAGR	Pixel acc	mAGR	Pixel acc	mAGR
[1]	61.0 ± 5.0	34.9 ± 3.9	49.9 ± 4.5	39.9 ± 4.8	49.5 ± 4.3	39.6 ± 4.4
Ours	74.5 ± 4.7	49.7 ± 3.8	68.5 ± 5.4	55.2 ± 4.5	68.4 ± 5.4	54.8 ± 4.3
Ours+Noc	77.4 ± 4.0	57.0 ± 3.3	70.2 ± 5.2	63.1 ± 4.5	70.1 ± 5.2	62.4 ± 4.3

H3D

Method	Full-a		Full-m		Unary	
	Pixel acc	mAGR	Pixel acc	mAGR	Pixel acc	mAGR
[1]	89.0 ± 0.8	63.4 ± 1.5	88.3 ± 0.8	69.6 ± 1.7	88.2 ± 0.8	69.8 ± 1.8
Ours	87.7 ± 0.8	62.7 ± 1.8	86.0 ± 1.0	70.2 ± 2.0	85.9 ± 1.1	70.4 ± 2.0

Fashionista

1 Yamaguchi et al, Parsing Clothes in Fashion Photographs.

Clothing Pattern Mining

Cloth Co-occurrence

- Upper-body and Lower-body cloth occurrences considered.
- Outermost cloth as the representative cloth label.
- Top results: Skirt-top, Shorts-top, Blouse-skirt, Tights-dress and Cardigan-dress

Clothing Pattern Mining

Color Co-occurrence

- Upper-body and Lower-body cloth color occurrences considered.
- Vector quantized dominant color of the representative cloth is color label.
- Top results (upper:lower): blue:blue, red:red, red:blue, blue:red, white:blue

Discussions

- Cloth parsing is formulated by incorporating more robust pose information for unrestricted settings.
- Proposed work outperforms previous state-of-the-art that uses Pictorial Structure based model.
- Qualitative observation shows that results are not reliable.
- The problem needs more breaking down. Even very accurate pose estimations can aid to an extent. More abstract information need to be parameterized like fine-grained cloth attributes, shape and layering order.

Related Publications

- Digvijay Singh, Vineeth Balasubramanian, and C V Jawahar. Fine-Tuning Human Pose Estimations in Videos. In WACV, 2016.

- Nataraj Jammalamadaka, Ayush Minocha, Digvijay Singh, and C V Jawahar. Parsing Clothes in Unrestricted Images. In BMVC, 2013.