Human pose estimation

Deya Ramanan
UC Irvine
Host of collaborators

Chaitanya Desai
Mohsen Hejrati
Hamed Pirsiavash
Dennis Park
Yi Yang
Xiangxin Zhu
David MxAllester
Pedro Felzenswzalb
Charless Fowlkes
Preface

This talk will contain a mix of “standard” tutorial material and “speculative” opinions

Please interrupt with questions!
Pattern classification versus visual understanding

Yes/no scanning window

Thursday, July 12, 2012
Pattern classification versus visual understanding

Yes/no scanning window

VS

“In-the-wild” pose estimation

Multiple bodies
Heavy occlusion
3D viewpoint
Pattern classification versus visual understanding

Yes/no scanning window

“In-the-wild” pose estimation

Multiple bodies
Heavy occlusion
3D viewpoint

use tools from here

Thursday, July 12, 2012
Why is finding people difficult?

variation in illumination

variation in appearance

variation in pose, viewpoint

occlusion & clutter

Classic “nuisance factors” for general object recognition
Appearance Templates

Rigid Templates

Quasi-rigid templates

Felzenszwalb, Girshick, McAllester, & Ramanan
PAMI 10
Why do parts help?

Star-models capture local affine (stretch, rotate, shear) deformations of template
Are quasi-rigid templates enough?
Deformable shape

Factored models of elastic geometry + appearance

Successful in graphics, but why not vision?

Too complex: inference is hard
A methodology of specifying a derivation under a rich visual grammar that we call visualization of the spatial models reflects the "cost" of placing the center of a part at different locations relative to the root.

To obtain high performance using discriminative training it is often important to use large training sets. In the case of object detection the training problem is highly unbalanced because there is vastly more background than objects. This motivates a process of searching through components of the score of that component model at the given location. In this case the latent information about the root is scored by a function $f_i, f_b, f_c$ of the following form:

$$z \in \Phi(x, \beta) = \max_{\Phi(x, \beta)}$$

where $z$ is a vector of model parameters, $x$ is a feature vector, $\beta$ is a specification of the object configuration, and $\Phi$ is a feature vector.

Fig 1. Detections obtained with a single component person model. The model is defined by a coarse root filter $f_{ia, f_b, f_c}$, several higher resolution part filters $f_{ib}$, and a spatial model for the location of each part relative to the root $f_{ic}$. The filters specify the Trifecta of shape $f_{ia, f_b, f_c}$.
Tractable shape

Star  Chain  Tree

Increasingly flexible
Overview

Background: part models

Articulation

Occlusion

3D viewpoint

Extensions
Old idea: part models

Model encodes local appearance + pairwise geometry
40 year history in vision

Pictorial Structures (Fischler & Elschlager 73, Felzenswalb and Huttenlocher 00)
Cardboard People (Yu et al 96)
Body Plans (Forsyth & Fleck 97)
Active Appearance Models (Cootes & Taylor 98)
Constellation Models (Burl et al 98, Fergus et al 03)
Background: deformable part models

\[ S(x, p) = \]

\[ x = \text{image} \]
\[ p_i = (x_i, y_i) \]
\[ p = \{z_1, z_2, \ldots\} \]
Background: deformable part models

\[ S(x, p) = \sum_{i} w_i \cdot \phi(x, p_i) + \text{part template scores} \]

\[ x = \text{image} \]
\[ p_i = (x_i, y_i) \]
\[ p = \{z_1, z_2, \ldots\} \]
Background: deformable part models

\[ S(x, p) = \sum_i w_i \cdot \phi(x, p_i) + \sum_{ij \in E} w_{ij} \cdot \psi(p_i, p_j) \]

- \( x \) = image
- \( p_i = (x_i, y_i) \)
- \( p = \{z_1, z_2, \ldots\} \)

\( \psi(p_i, p_j) = [dx \quad dx^2 \quad dy \quad dy^2]^T \)

\( E = \) relational graph
Shape term

\[ \sum_{ij \in E} w_{ij} \cdot \psi(p_i, p_j) \]

Can make pairwise term image-independant \( \psi(x, p_i, p_j) \)

Sapp et al ECCV,CVPR,NIPS 2010
Tran & Forsyth ECCV 10
Shape term

\[
\sum_{i,j \in E} w_{ij} \cdot \psi(p_i, p_j) = (p - \mu)^T \Lambda (p - \mu)
\]

where \((\mu, \Lambda)\) are functions/reparameterizations of \(\{w_{ij}\}\)
and \(\Lambda\) is the block-sparse inverse of a shape “covariance” matrix.
Shape term

\[ \sum_{ij \in E} w_{ij} \cdot \psi(p_i, p_j) = (p - \mu)^T \Lambda (p - \mu) \]

where \((\mu, \Lambda)\) are functions/reparameterizations of \(\{w_{ij}\}\)
and \(\Lambda\) is the block-sparse inverse of a shape “covariance” matrix

Lesson: stars don’t deform that much, but trees do!
Shape term (derivation)

\[
\sum_{ij \in E} a_{ij} dx^2 + b_{ij} dx + c_{ij} dy + d_{ij} dy^2 = \sum_{ij \in E} \left( \begin{pmatrix} p_i - \mu_i \\ p_j - \mu_j \end{pmatrix} \right)^T \Lambda_{i,j} \left( \begin{pmatrix} p_i - \mu_i \\ p_j - \mu_j \end{pmatrix} \right) + \text{constant}, \quad \text{where} \quad \Lambda_{i,j} = -\begin{bmatrix}
    a_{ij} & 0 & -a_{ij} & 0 \\
    0 & c_{ij} & 0 & -c_{ij} \\
    -a_{ij} & 0 & a_{ij} & 0 \\
    0 & -c_{ij} & 0 & c_{ij}
\end{bmatrix}
\]
Inference: \( \max_{p} S(x,p) \)

- \( N \) candidate locations, \( K \) parts
- Dynamic programming reduces search from \( O(N^K) \) to \( O(KN^2) \) for trees
- For each candidate head, independently estimate best left and right arm
- In practice, no more expensive than scoring each part independently

Felzenszwalb & Huttenlocher IJCV 05

Markov model
Inference: $\max_{p} S(x,p)$

1) Initialize nodes with match score
2) Initialize edges with spring score
3) Find best path from left to right

In practice, (1) is bottleneck
Our total performance of 89.6 fl.s compares favorably to the best previous result of 82.4 fl.s. We also beat all previous results.

Figure 7: Visualization of our full body model for mixture models for the hand are spread apart to capture a larger variety of relative orientations. These clusters are used to generate mixture labels for parts during training. For example, heads tend to be upright and so the associated mixture models focus on upright orientations.

Figure 5: We take a “data-driven” approach to orientation modeling by clustering the relative locations of parts with respect to their parents. These clusters are used to generate mixture labels for parts during training.

Score is linear in local templates $w_i$ and spring parameters $w_{ij}$
Learning linear parameters

\[ f_w(x) = w \cdot \Phi(x) \]

Training data consists of images with labeled bounding boxes

- Need to learn the model structure, filters and deformation costs

- Positive weights
- Negative weights

Train ‘w’ with linear classifier (perceptron, SVM, regression, …)
Learning linear parameters

\[ f_w(x) = w \cdot \Phi(x) \]

- Training data consists of images with labeled bounding boxes
- Need to learn the model structure, filters and deformation costs

\[ \min_w \|w\|^2 \]

\[ \forall i \in \text{pos}, \ w \cdot x_i > 1 \]

\[ \forall i \in \text{neg}, \ w \cdot x_i < 1 \]
Large-scale learning

Our test set distribution is highly imbalanced; so should be the training set
(hundreds of positives, hundreds of millions of negatives)

SVMs are attractive because they generate sparse learning problems
(One can solve problems that are too big to fit in memory)
Learning structured linear parameters

\[ S(x, p) = w \cdot \Phi(x, p) \]

(pos)

(neg)

(Apply same sparse learning tricks to deal with exponential set of negatives!)

Thursday, July 12, 2012
Learning structured linear parameters

\[ S(x, p) = w \cdot \Phi(x, p) \]

(Apply same sparse learning tricks to deal with exponential set of negatives!)
Perhaps we don’t even need SVMs?

Learn templates with simple statistical Gaussian models

\[ w = \Sigma^{-1}(\mu_1 - \mu_0) \]

Hariharan, Malik, Ramanan ECCV 12
Datasets

H3D Berkeley

Buffy Oxford

Leeds Sports dataset

PASCAL Stickman
PASCAL Layout Competition

(the forgotten challenge)

Makes use of DPM to detect candidate parts & reranks with SVM

<table>
<thead>
<tr>
<th></th>
<th>head</th>
<th>hand</th>
<th>foot</th>
</tr>
</thead>
<tbody>
<tr>
<td>OXFORD_RANK_SLIACK_RBF</td>
<td>72.9</td>
<td>26.9</td>
<td>4.1</td>
</tr>
</tbody>
</table>
Overview

Background: part models

Representations

Occlusion

3D variation

Extensions
What’s wrong with part models?

(Flawed) assumption: local appearance and global geometry are independent

(e.g., head looks the same no matter the geometry of the rest of the body)

When does this fail?
What’s wrong with part models?

(Flawed) assumption: local appearance and global geometry are independent

(e.g., head looks the same no matter the geometry of the rest of the body)

When does this fail?

Articulation 3D viewpoint Occlusion
Modeling articulation

Enlarge state space of part location to include orientation and foreshortening

\[(x_i, y_i) \Rightarrow (x_i, y_i, \theta_i, s_i)\]

Problem: rather expensive and doesn’t work well
One solution: local mixtures of small patches
Any smooth spatial transformation is locally rigid
Local mixtures of parts

\[
S(x, p, t) = \sum_i w_i^{t_i} \cdot \phi(x, p_i) + \sum_{ij \in E} w_{ij}^{t_i, t_j} \cdot \psi(p_i, p_j)
\]

Each part has a position ‘p’ and mixture type ‘t’

Score local model with one of T templates

Score deformation with one of $T^2$ springs (interdependence of geometry + appearance)
Local mixtures of parts

Each part has a position ‘p’ and mixture type ‘t’

Score local model with one of T templates

Score deformation with one of T² springs (interdependence of geometry + appearance)

\[
S(x, p, t) = \sum_i w_i^t \cdot \phi(x, p_i) + \sum_{ij \in E} w_{ij}^{t_i, t_j} \cdot \psi(p_i, p_j) + S(t)
\]

\[
p_i = (x_i, y_i)
\]

\[
t_i \in \{1, \ldots, T\}
\]
Appearance relations

Spring co-occurrence prior

\[ S(t) = \sum_{ij \in E} b_{ij}^{t_i,t_j} \]

Rigid relation

Flexible relation
Supervised learning

\[ S(x, p, t) = w \cdot \Phi(x, p, t) \]

Given \( \{x_n, p_n, t_n\} \), tune ‘w’ such that \( S(x, p, t) \)
scores high on people and low on backgrounds
(structured prediction)
Inference

Consider “joint” domain of part location and mixture type: \( z_i = (p_i, t_i) \)

\[
S(z) = \sum_i \phi_i(z_i) + \sum_{ij \in E} \psi_{ij}(z_i, z_j)
\]

(simple discrete tree-MRF)

Pixel locations and mixture types

head  torso  leg
Exponential number of global mixtures

K parts, M local mixtures ⇒ $K^M$ unique global mixtures

Not all combinations are equally likely; “prior” given by co-occurrence model
Qualitative Results

Yi & Ramanan CVPR11
Search over representations

Denser parts and more local mixtures help (up to a point)

---

14 parts

26 parts

---

Performance vs # of parts (K) and mixtures (T)

- 14-part model
- 26-part model
- 51-part model

Number of mixtures (T) vs Probability of correct keypoints (PCK)
Quantitative evaluation

% of correctly localized limbs

<table>
<thead>
<tr>
<th>Method</th>
<th>Torso</th>
<th>Head</th>
<th>Upper legs</th>
<th>Lower legs</th>
<th>Upper arms</th>
<th>Lower arms</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>R [23]</td>
<td>52.1</td>
<td>37.5</td>
<td>31.0</td>
<td>29.0</td>
<td>17.5</td>
<td>13.6</td>
<td>27.2</td>
</tr>
<tr>
<td>ARS [1]</td>
<td>81.4</td>
<td>75.6</td>
<td>63.2</td>
<td>55.1</td>
<td>47.6</td>
<td>31.7</td>
<td>55.2</td>
</tr>
<tr>
<td>JEa [15]</td>
<td>77.6</td>
<td>68.8</td>
<td>61.5</td>
<td>54.9</td>
<td>53.2</td>
<td>39.3</td>
<td>56.4</td>
</tr>
<tr>
<td>SNH [29]</td>
<td>91.2</td>
<td>76.6</td>
<td>71.5</td>
<td>64.9</td>
<td>50.0</td>
<td>34.2</td>
<td>60.9</td>
</tr>
<tr>
<td>JEb [14]</td>
<td>85.4</td>
<td>76.1</td>
<td>73.4</td>
<td>65.4</td>
<td>64.7</td>
<td>46.9</td>
<td>66.2</td>
</tr>
<tr>
<td>Our Model</td>
<td><strong>97.6</strong></td>
<td><strong>93.2</strong></td>
<td><strong>83.9</strong></td>
<td><strong>75.1</strong></td>
<td><strong>72.0</strong></td>
<td><strong>48.3</strong></td>
<td><strong>74.9</strong></td>
</tr>
</tbody>
</table>

On-par with or outperforms previous work while being orders of magnitude faster
(few seconds vs few minutes)

All previous work use explicitly articulated models
Model affine warps of templates with mixtures of pictorial structures

Faster run-time
(small templates + dynamic programming)
What makes it work better?

<table>
<thead>
<tr>
<th></th>
<th>Joint</th>
<th>Indep</th>
<th>Indep+Invar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>67.4</td>
<td>51.3</td>
<td>33.8</td>
</tr>
</tbody>
</table>
Why does joint training help?

We need compete only against joint configurations of negatives that score above margin
Why are parts not orientation-invariant?

<table>
<thead>
<tr>
<th>Joint</th>
<th>Indep</th>
<th>Indep+Invar</th>
</tr>
</thead>
<tbody>
<tr>
<td>67.4</td>
<td>51.3</td>
<td>33.8</td>
</tr>
</tbody>
</table>

Illumination (world is lit from above)
Occlusions (torsos tend to be upright)
Overview

Background: part models

Articulation

Occlusion

3D variation

Extensions
Representations for human pose

- **Patches**: Smaller parts
- **Skeleton**: Larger parts
- **Poselets**: Smaller parts

Thursday, July 12, 2012
Representations for human pose

Skeleton

Larger parts

Poselets
Global representations

Skeleton
Ioffe & Forsyth
zenswalb & Huttenlocher
ohnson & Everingham
Andruikula et al.
Ferrari et al.

Poselets
Bourdev & Malik
Maji et al.
Yang & Mori
Wang & Yang

Exemplars
Malisiewicz et al
Mori & Malik
Shaknarovich & Darrell
Johnson & Everingham

Visual Phrases
Sadeghi and Fahardi
Global representations

Skeleton  Poselets  Exemplars  Visual Phrases

Insight from such global approaches (an opinion):
large composite templates better model occlusions and interactions
How to encode complex interactions?

One may need lots of large composite templates

Visual Phrases
Sadeghi and Fahardi, CVPR 11
How to encode complex interactions?

Poselets
Bourdev & Malik ICCV09

One may need lots of large composite templates
One take: visual “phraselets”

Person on horse

Person on jumping horse

Person standing next to horse

Break up visual composite into smaller patches and reason about appearance relations
One take: visual “phraselets”

Hand looks different due to interactions with global geometry

We’ll encode such visual differences as local part mixtures
Learning phraselets

Define phraselets as commonly-occurring geometric configurations

“Poselet-like clusters”

Given labelled training data, find clusters of keypoint configurations relative to each joint
this sections we describe a method for obtaining mixture labels. Our intuition is that from an activity with keypoint labels spanning both the human body and any different semantic object types; for example, head, lt shoulder, rt shoulder, elbow, lt ankle, etc. We assume these keypoint labels are with a visibility flag denoting if a particular keypoint is occluded or not.

In Sec 5, we write as described in section ws, our clusters naturally differ from interactions. Visual phrases [5] takes a "bruteforce" approach to model a single instance of a person/object in the image. Our work differs from this in that the bottom left, center, and bottom right clusters encode similar viewpoints, but differ in that the bottom left cluster encodes changes in viewpoint, as well as different semantic object types; for example, front wheel, rear wheel, bike handle.

We describe our approach for learning phraselets or mixtures of local patches. This approach may require a separate template for the person and object. This approach may require a separate template for both the person and object. This approach may require a separate template for both the person and object. This approach may require a separate template for both the person and object. This approach may require a separate template for both the person and object.

For the central figure and bottom center and bottom right clusters, the encoding of similar viewpoints, but different viewpoints, may require a separate template for different viewpoints. This approach may require a separate template for different viewpoints. This approach may require a separate template for different viewpoints. This approach may require a separate template for different viewpoints. This approach may require a separate template for different viewpoints. This approach may require a separate template for different viewpoints. This approach may require a separate template for different viewpoints.

For a mixture or phraselet label, we show bike handles from PASCAL "riding bike" action clustered using different viewpoints. The clusters naturally differ from interactions. Visual phrases [5] takes a "bruteforce" approach to model a single instance of a person/object in the image. Our work differs from this in that the bottom left, center, and bottom right clusters encode similar viewpoints, but differ in that the bottom left cluster encodes changes in viewpoint, as well as different semantic object types; for example, front wheel, rear wheel, bike handle.
Model occlusions with separate clusters

Visible left elbow

Occluded left elbow

Mixture label corresponds to visible/occlusion state
Local mixtures of phraselets

\[ S(x, p, t) = \sum_i w_i^{t_i} \cdot \phi(x, p_i) + \sum_{ij \in E} w_{ij}^{t_i, t_j} \cdot \psi(p_i, p_j) + S(t) \]

Relational model encodes that one (but not both) the left & right leg is occluded when they are nearby.
Relational phraselets

Report back human+object part locations and mixture label

Desai and Ramanan ECCV 12
Fig. 5. We show detection results obtained without any manual annotation. We follow the notational conventions of Fig. 1, including open circles to denote occluded parts. Note that our compositional models are able to capture large changes in viewpoint and articulation that are present even within a single action class.
Results

Detecting person-object interactions

Red line: Our compositional phraselet model
Blue line: DPM trained on person+object (visual phrase)
Top false positives

(penalized detections)

Using computer  Taking photo
Action classification

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Us</td>
<td>81.6</td>
<td>82.6</td>
<td>84.7</td>
<td>52.6</td>
<td>25.3</td>
<td>54.9</td>
<td>56.4</td>
</tr>
<tr>
<td>Poselets</td>
<td><strong>83.1</strong></td>
<td>79.9</td>
<td><strong>87.6</strong></td>
<td>45.9</td>
<td><strong>26.2</strong></td>
<td>44.9</td>
<td><strong>66.6</strong></td>
</tr>
</tbody>
</table>

For action classification (given known bounding box), method near state-of-art
Person-person composites

Yang et al. “Recognizing Proxemics in Personal Photo Collections” CVPR12
Proxemic analysis

Edward Hall  “A system for the notation of proxemic behavior”
American Anthropologist 1963

Relative body orientation

Touching body parts
(heads, elbows, hands)
Multi-body pose estimation

Eichner & Ferrari. “We are Family: Joint Pose Estimation” ECCV 2010

Yang et al. “Recognizing Proxemtics in Personal Photo Collections” CVPR12
### Dataset statistics

<table>
<thead>
<tr>
<th>(a) Image Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Images</td>
<td>No. People</td>
</tr>
<tr>
<td>589</td>
<td>1207</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) Touch Code Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand-hand</td>
<td>Hand-shoul</td>
</tr>
<tr>
<td>340</td>
<td>180</td>
</tr>
<tr>
<td>25.5%</td>
<td>13.5 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) Co-occurrence Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Codes</td>
<td>1 Code</td>
</tr>
<tr>
<td>531</td>
<td>626</td>
</tr>
</tbody>
</table>
Quantitative results

Sequential approach: (1) Estimate pose with single-body model
(2) Classify touch-code based on estimate pose

Yang et al. “Recognizing Proxemics in Personal Photo Collections” CVPR12
Quantitative results

Removing key spring (yellow) drops performance from 52% to 33%

Correct spatial structure is crucial
(e.g., difficult to reason about with a star model)
Overview

Background: part models

Occlusion reasoning

3D variation

Extensions
View-based models of faces

Ioffe & Forsyth, 2001
Everingham, Sivic, & Zisserman, 2006

Use global mixtures to capture topological changes due to viewpoint
Use common pool of parts
View-based models of faces

Ioffe & Forsyth, 2001
Everingham, Sivic, & Zisserman, 2006

Use **global** mixtures to capture topological changes due to viewpoint
Use common pool of parts
View-based models

Model self-occlusion with missing branches
View-based models

Model self-occlusion with missing branches
Learning

Fully-supervised dataset (CMU MultiPIE)

Chow-Liu algorithm
Global models of deformation

Full-covariance Gaussian shape

Tree-based max-margin shape

Thursday, July 12, 2012
Learned appearance & deformation
Global mixtures capture large viewpoint changes

Elastic springs capture small viewpoint changes

... all without explicit 3D reasoning
Evaluation on Flickr images
Qualitative results

Model simultaneously addresses face detection, pose estimation, and landmark localization
Winner of LFW face recognition challenge
Y. Taigman, L. Wolf, 2011
Detection results

Dalal & Triggs, 2005

Detection results

- Multiview HoG
- Google Picasa
- face.com
- 2-view Viola Jones
Detection results

Dalal & Triggs, 2005

Multiview HoG
Google Picasa
face.com
2–view Viola Jones

Thursday, July 12, 2012
Detection results

Felzenszwalb et al. 2010

- DPM, voc-re4
- Multiview HoG
- Google Picasa
- face.com
- 2-view Viola Jones

Precision vs. Recall
Detection results

Felzenszwalb et al. 2010

- DPM, voc-re4
- Multiview HoG
- Google Picasa
- face.com
- 2-view Viola Jones

Def. Parts
Multiview
Detection results

Our star model
DPM, voc-re4
Multiview HoG
Google Picasa
face.com
2-view Viola Jones

Def. Parts
Multiview

Thursday, July 12, 2012
Detection results

- Our star model
- DPM, voc-re4
- Multiview HoG
- Google Picasa
- face.com
- 2-view Viola Jones

Detection results graph with precision on the y-axis and recall on the x-axis.
Detection results

- Our model
- Our star model
- DPM, voc-re4
- Multiview HoG
- Google Picasa
- face.com
- 2-view Viola Jones

Thursday, July 12, 2012
Landmark localization

Our model is learned with only hundreds of faces

vs

Taigman & Wolf “Leveraging Billions of Faces...” 2011
Landmark localization

Baselines are initialized with ground truth detection on test images.
Our model naturally produces state-of-the-art pose and landmark estimates

Thursday, July 12, 2012
A look back: why do part models help?

Consider a $K$-part model, with $L$ discrete part locations.

At run-time, part model = exponentially-large $O(L^K)$ mixture of rigid templates.
A look back: why do part models help?

Consider a K-part model, with L discrete part locations.

At run-time, part model = exponentially-large \( O(L^K) \) mixture of rigid templates.

Compared to a mixture of exemplars (Malisiewicz et al), part models...

1) Share parameters across mixtures
2) “Synthesize” new rigid templates not seen during training
3) Efficiently search over mixtures using dynamic programming

Thursday, July 12, 2012
A look back: why do part models help?

1) Share parameters across mixtures
2) “Synthesize” new rigid templates not seen during training

To examine (1) vs (2), let's define mixture of exemplars with sharing
An analysis of part models

Zhu et al, BMVC 2012
An analysis of part models

Zhu et al, BMVC 2012
An analysis of part models

Zhu et al, BMVC 2012

“hallucinate” new mixtures by composing parts

Num. of training samples

Average precision

Sup.DPM

Mix. with tying

Mix. HoG templ.
An analysis of part models

Zhu et al, BMVC 2012
Overview

Background: part models

Occlusion reasoning

3D variation

Extensions
Challenges in scalability:

Vocabularies of thousands of parts

Is there a more efficient representation?
Steerable basis

Freeman, Adelson, Perona

\[ w_i = \sum_j s_{ij} b_j \]

\[ \approx \text{linear combinations of basis templates} \]

This can be implemented as a rank-restriction on original set of templates
Learning steerable part models

Learn vocabularies of thousands of parts

\[ w_i = \sum_j s_{ij} b_j \]

Learn rank-constrained linear classifiers with off-the-shelf structural SVM solvers
Steerable (& separable) part models

Pirsiavash & Ramanan CVPR12

Models are 10-100X smaller & faster with near-equivalent performance

Share “soft” basis rather than fixed templates (across views/categories)

Philosophy: We should treat parameters \( w \) as spatial filters, not vectors
Non-tree constraints: occlusion

How to handle “loopy” constraints that arise from occlusion phenomena?

Sigal & Black CVPR 06
Non-tree constraints: appearance

Pairwise consistency (symmetry in appearance)

Global consistency (latent appearance)
Tools for inference on non-trees

One approach: apply standard approximate inference algorithms for Markov Random Fields (MRFs)

Why is this hard?

1) Large discrete domains of variables (e.g., pixels in an image)

2) Continuous domains of variables (e.g., color and appearance)
Tools for inference on non-trees

One successful approach: use tree-like inference algorithms

**Mixtures of trees** (condition on mixture variable)
Ioffe & Forsyth, Johnson & Everginham, Lan & Huttenlocher, Wang & Mori

**Loopy Belief Propagation** (iteratively apply tree-based messages)
Sigal and Black

**Dual Decomposition** (break problem up into trees ensuring agreement)
Sapp et al, Kumar et al

**Branch & Bound** (use trees to generate strong lower bounds)
Tian and Scarloff, Nevatia

**Sampling** (importance sample from tree)
Felzenszwalb & Huttenlocher, Beuhler et al
N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex (loopy) model

Popular in speech, but why not vision?
N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex (loopy) model

Popular in speech, but why not vision?
N-best maximal decoding

Use max-marginals + NMS to compute the “next-best non-overlapping pose”

Park and Ramanan, ICCV11
Yadollahpour et al. ECCV12
N-best maximal decoding

Intuition: backtrack from all parts, not just root
(can we done without any noticeable increase in computation)

Park and Ramanan, ICCV 2011
N-best maximal decoding

Philosophy: Delay hard decisions as much as possible

Candidate interest points
Candidate parts
Candidate poses
Maximal poses from a single frame

Correct one picked out by temporal context (tracker)
Evaluation

Outperforms standard approaches by 20%
Just as fast as finding single-best configuration

Percentage of correct frames

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>walking</th>
<th>pitching</th>
<th>lola1</th>
<th>lola2</th>
</tr>
</thead>
<tbody>
<tr>
<td>noNMS</td>
<td>0.825</td>
<td>0.762</td>
<td>0.505</td>
<td>0.445</td>
</tr>
<tr>
<td>rootNMS</td>
<td>0.815</td>
<td>0.741</td>
<td>0.455</td>
<td>0.390</td>
</tr>
<tr>
<td>partNMS</td>
<td>0.825</td>
<td>0.762</td>
<td>0.515</td>
<td>0.420</td>
</tr>
<tr>
<td>MMsmpl</td>
<td>0.930</td>
<td>0.800</td>
<td>0.645</td>
<td>0.440</td>
</tr>
<tr>
<td>Nbest(all)</td>
<td>0.940</td>
<td><strong>0.800</strong></td>
<td>0.635</td>
<td>0.495</td>
</tr>
<tr>
<td>Nbest(limb)</td>
<td><strong>0.950</strong></td>
<td>0.797</td>
<td>0.670</td>
<td>0.500</td>
</tr>
</tbody>
</table>
A look back

Articulation

Visual composites

3D aspect

Underlying theme: tractable, joint representations of shape and appearance
Thank you!