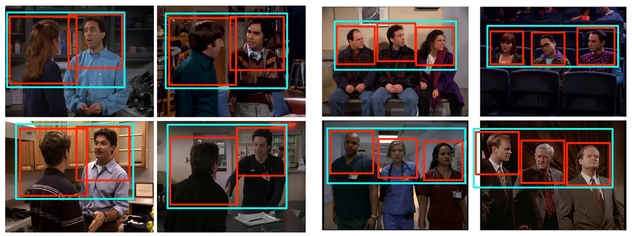
Talking Heads: Detecting Humans and Recognizing Their Interactions

Minh Hoai, Andrew Zisserman
Visual Geometry Group, University of Oxford

Objective and Key Idea

- Humans are the primary focus of many TV shows. Detecting them is crucial for understanding TV material.
- We propose an algorithm for detecting people by reasoning about their common configurations in TV shows.
- Observe the similarity of the following upper body configurations:

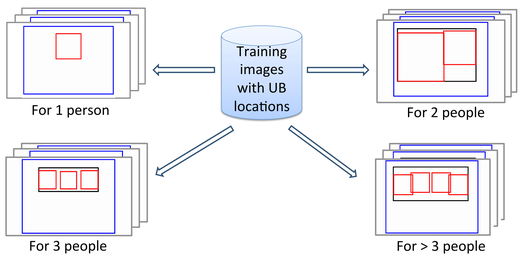


A common configuration for 2 people

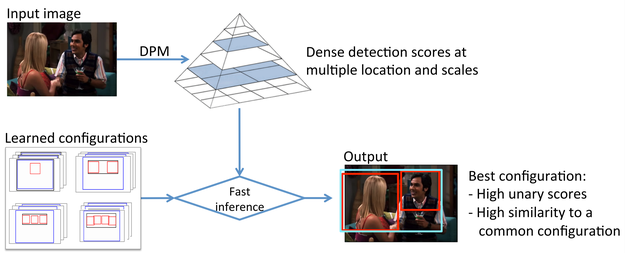
A common configuration for 3 people

Approach Overview

Learning the common configurations for 1, 2, 3, and > 3 people

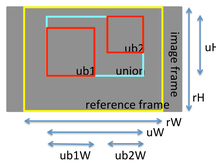


Detection Procedure



Details

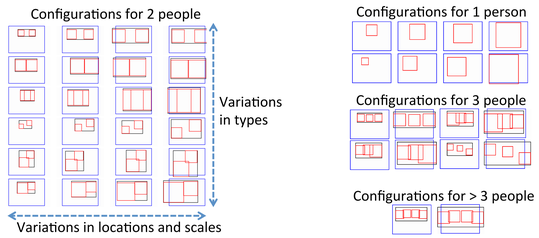
Quantifying a configuration of upper bodies:



- Construct two configuration vectors
- Level-1 configuration:
  ub1 and ub2 w.r.t. to the union
- Level-2 configuration:
  The union w.r.t. the reference frame
- Configuration:
  Relative locations and scales, e.g., for level-2:

$$\left[\frac{uX - rX}{rW}, \frac{uY - rY}{rH}, \log\left(\frac{uW}{rW}\right), \log\left(\frac{uH}{rH}\right)\right]$$

Learning Common configurations with Hierarchical Clustering



- Many configurations drawn above have a left-right mirror version
- Total number of learned configurations for 1, 2, 3, and > 3 UBs are: 12, 36, 10, and 2, respectively

Energy and Inference

Energy: $E(\{\mathbf{p}_i\}, \mathbf{u}) = \min_{\theta \in \Theta} E(\{\mathbf{p}_i\}, \mathbf{u}|\theta)$

A set of UBs UB union — A configuration model

$$E(\{\mathbf{p}_i\}, \mathbf{u}|\theta) = \sum_{i=1}^{k} \alpha_i \mathcal{U}(\mathbf{p}_i) + \sum_{i=1}^{k} \boldsymbol{\beta}_i^T \phi_1(\mathbf{p}_i|\mathbf{u}) + \boldsymbol{\gamma}^T \phi_2(\mathbf{u}) + b$$

Unary potential — Relative scale and location b/t UB and the UB union — Relative scale and location b/t the UB union and the image

Parameters of configuration model

Inference is fast:
- The dependency between variables is a tree structure. This enables dynamic programming.
- Generalized Distance transform can be used.
- Much computation can be shared between configuration models

Learning the parameters of configuration models

Assume we have labeled training data $\{\mathbf{I}_i, \mathbf{P}_i, y_i\}$ (image, UBs, configuration model)

Energy function is linear in parameters, rewrite $E_k(\mathbf{I}, \mathbf{P}) = -(\mathbf{w}_k^T \varphi_k(\mathbf{I}, \mathbf{P}) + b_k)$

Parameters to learn — Unary potentials + deformation

Max-margin learning:

$$\underset{\{\mathbf{w}_k, b_k, \xi_i \geq 0\}}{\text{minimize}} \ \frac{1}{2m} \sum_{1}^{m} ||\mathbf{w}_k||^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } \mathbf{w}_{y_i}^T \varphi_{y_i}(\mathbf{I}_i, \mathbf{P}_i) + b_{y_i} \geq \mathbf{w}_y^T \varphi_y(\mathbf{I}_i, \mathbf{P}) + b_y + 1 - \xi_i$$
$$\forall i, \mathbf{P}, y : n_y \neq n_{y_i}.$$

Experiments

Datasets

- TV Human Interaction (TVHI): 300 video clips from 23 different TV shows
- Our Sitcom Dataset: frames extracted from 150 episodes of The Big Bang Theory, Scrubs, Seinfeld, and Frasier

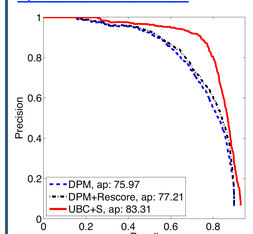
Number of UBs	0	1	2	3	≥4	total
TVHI train data	0	118	370	79	32	599
TVHI test data	0	100	464	121	29	714
Combined train data	143	448	740	291	32	1654
Combined test data	128	406	726	227	29	1566

Detection Examples

Failure cases

UBC + a Singleton Detector (UBC + S)

Quantitative Evaluation



UBC still outperforms DPM even with tracking enhancement.
E.g., Precision at 80% recall:
DPM: 74.45%
UBC: 84.18%

Additional Experiments

Sensitivity Analysis

No. of 1-UB CMs	8	8	12	20	20
No. of 2-UB CMs	24	40	40	40	64
AP on TVHI data.	81.33	81.58	81.93	81.89	82.02
AP on Com. data.	85.39	85.96	86.56	86.58	86.83

Average Precision (AP) as the number of configuration models varies. UBC is not too sensitive to this setting.

Upper-body counting – Confusion matrices

DPM

Actual \	0	1	2	3	≥4
0	.98	.02	.00	.00	.00
1	.12	.67	.21	.00	.00
2	.11	.31	.41	.14	.02
3	.04	.10	.29	.36	.21
≥4	.00	.22	.21	.34	.22

Acc: 52.84%

UBC+S (ours)

Actual \	0	1	2	3	≥4
0	.95	.02	.02	.01	.00
1	.05	.87	.07	.00	.00
2	.04	.21	.67	.07	.01
3	.02	.03	.36	.53	.06
≥4	.00	.00	.33	.33	.34

Acc: 67.09%

Human Interaction Recognition

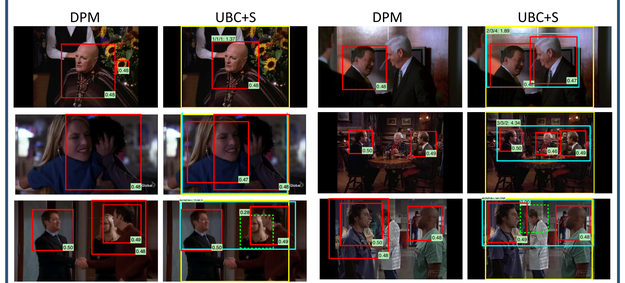
	Handshake	Highfive	Hug	Kiss	Mean
Patron *et al.* [16]	39.4	45.8	47.0	37.6	42.4
Marin *et al.* [14]	-	-	-	-	39.2
Yu *et al.* [29]	-	-	-	-	55.9
Gaidon *et al.* [9]	-	-	-	-	55.6
DTD [9, 25]	-	-	-	-	53.4
Ours	**55.8**	**60.2**	**60.8**	**48.2**	**56.3**

Average Precision on TVHI dataset.
We use Dense Trajectory Descriptors

Detection Speed

On Matlab 2.3 GHz CPU, for a 352x624 image:
- Computing dense scores (using DPM): 945ms
- Additional UBC inference: 610ms for 60 models

More detection examples



Code available: www.robots.ox.ac.uk/~vgg/software/ubc/