All you want to know about GPs: Applications of GPs

Raquel Urtasun and Neil Lawrence

TTI Chicago, University of Sheffield

June 16, 2012

Applications of Gaussian Process Regression

We will concentrate on a few successful applications in computer vision

1. Multiple kernel learning: object recognition
2. Active Learning: object recognition
3. GPs as an optimization tool: weakly supervised segmentation
4. Human pose estimation from single images
5. Flow Classification and ROI detection
6. Shape from shading
7. Online Shopping

1) Object Recognition

- **Task:** Given an image $\mathbf{x}$, predict the class of the object present in the image $\mathbf{y} \in \mathcal{Y}$

$$y \rightarrow \{car, bus, bicycle\}$$

- Although this is a classification task, one can treat the categories as real values and formulate the problem as regression.

1) Object Recognition

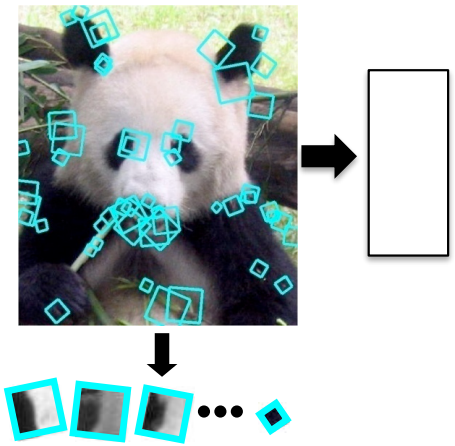
- **Task:** Given an image $\mathbf{x}$, predict the class of the object present in the image $\mathbf{y} \in \mathcal{Y}$

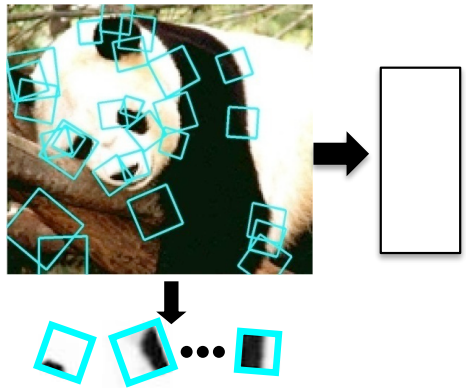
$$y \rightarrow \{car, bus, bicycle\}$$

- Although this is a classification task, one can treat the categories as real values and formulate the problem as regression.

How do we do Object Recognition?

- Given this two images, we will like to say if they are of the same class.





- Choose a representation for the images
  - Global descriptor of the full image
  - Local features: SIFT, SURF, etc.
- We need to choose a way to compute similarities
  - Histograms of local features (i.e., bags of words), pyramids, etc.
  - Kernels on global descriptors, e.g., RBF
  - ...

Multiple Kernel Learning (MKL)

- Why do we need to choose a single representation and a single similarity function?
- Which one is the best among all possible ones?
- Multiple kernel learning comes at our rescue, by learning which cues and similarities are more important for the prediction task.
- Simplest form:

$$\mathbf{K} = \sum_i \alpha_i \mathbf{K}_i$$

Multiple Kernel Learning (MKL)

- Why do we need to choose a single representation and a single similarity function?
- Which one is the best among all possible ones?
- Multiple kernel learning comes at our rescue, by learning which cues and similarities are more important for the prediction task.
- Simplest form:

$$\mathbf{K} = \sum_i \alpha_i \mathbf{K}_i$$

- This is just hyperparameter learning in GPs! No need for specialized SW!

Multiple Kernel Learning (MKL)

- Why do we need to choose a single representation and a single similarity function?
- Which one is the best among all possible ones?
- Multiple kernel learning comes at our rescue, by learning which cues and similarities are more important for the prediction task.
- Simplest form:

$$\mathbf{K} = \sum_i \alpha_i \mathbf{K}_i$$

- This is just hyperparameter learning in GPs! No need for specialized SW!

Efficient Learning Using GPs for Multiclass Problems

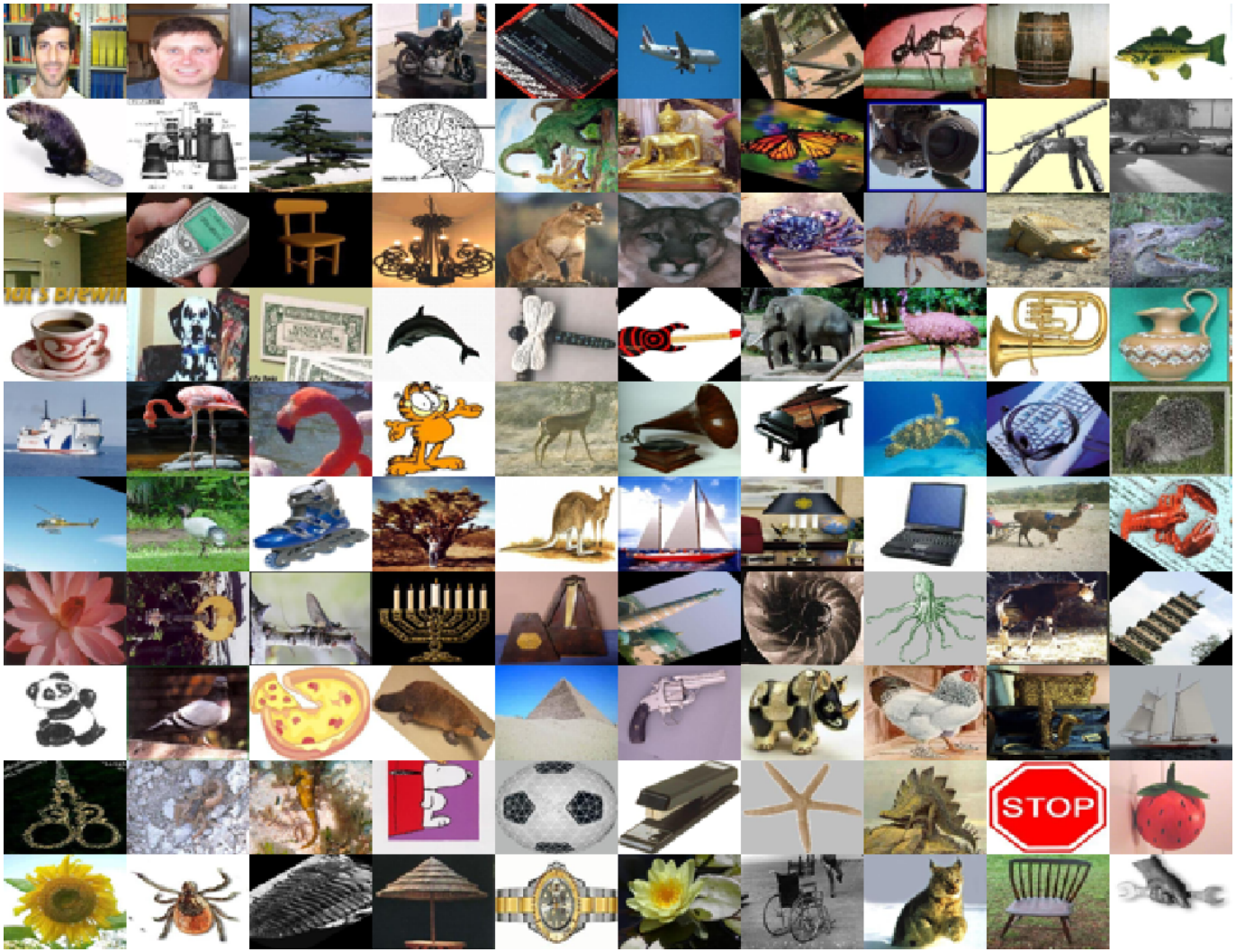
Supposed we want to emulate a 1-vs-all strategy as $|\mathcal{Y}| > 2$

- We define $\mathbf{y} \in \{-1, 1\}^{|\mathcal{Y}|}$
- We can employ maximum likelihood and learn all the parameters for all classifiers at once

$$\min_{\boldsymbol{\theta}, \boldsymbol{\alpha} > 0} -\sum_i \log p(\mathbf{y}^{(i)} | \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\alpha}) + \gamma_1 ||\boldsymbol{\alpha}||_1 + \gamma_2 ||\boldsymbol{\alpha}||_2$$

  with $\mathbf{y}^{(i)} \in \{-1, 1\}$ each of the individual problems.
- Efficient as we can share the covariance across all classes

Caltech 101 dataset



Results: Caltech 101

[A. Kapoor, K. Graumann, R. Urtasun and T. Darrell, IJCV 2009]

Comparison with SVM kernel combination: kernels based on Geometric Blur (with and without distortion), dense PMK and spatial PMK on SIFT, etc.

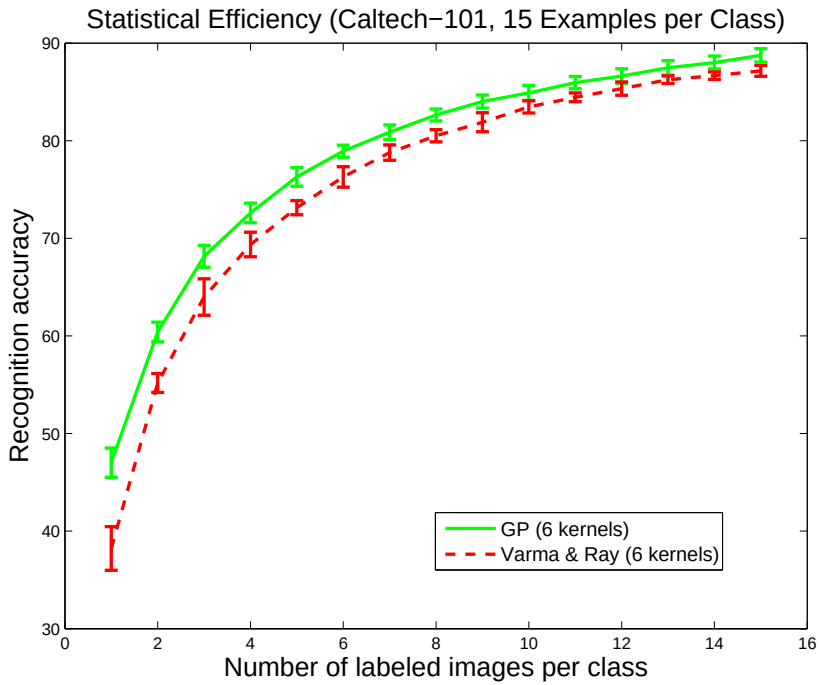


Figure: Average precision.

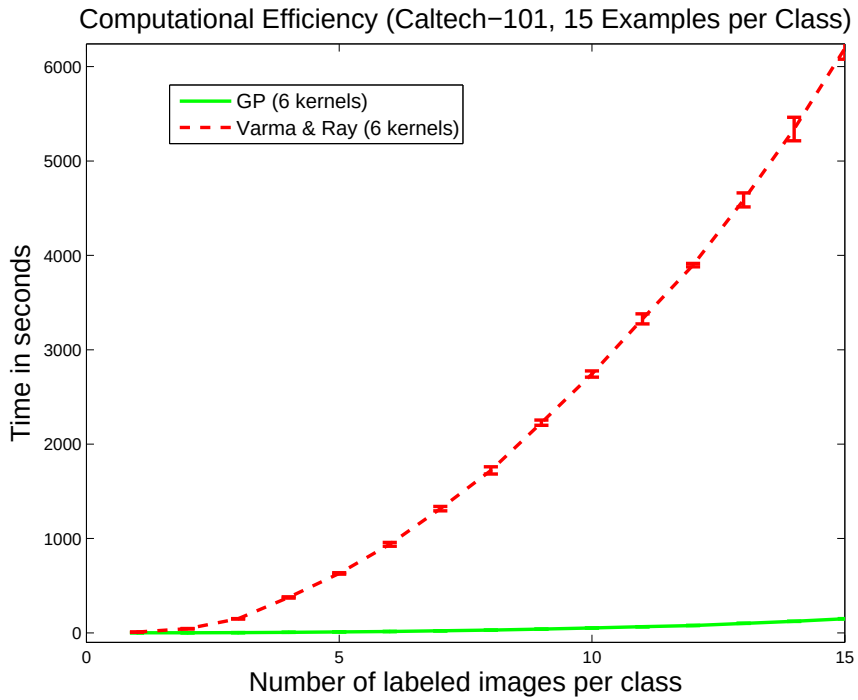


Figure: Time of computation.

Results: Caltech 101

[A. Kapoor, K. Graumann, R. Urtasun and T. Darrell, IJCV 2009]

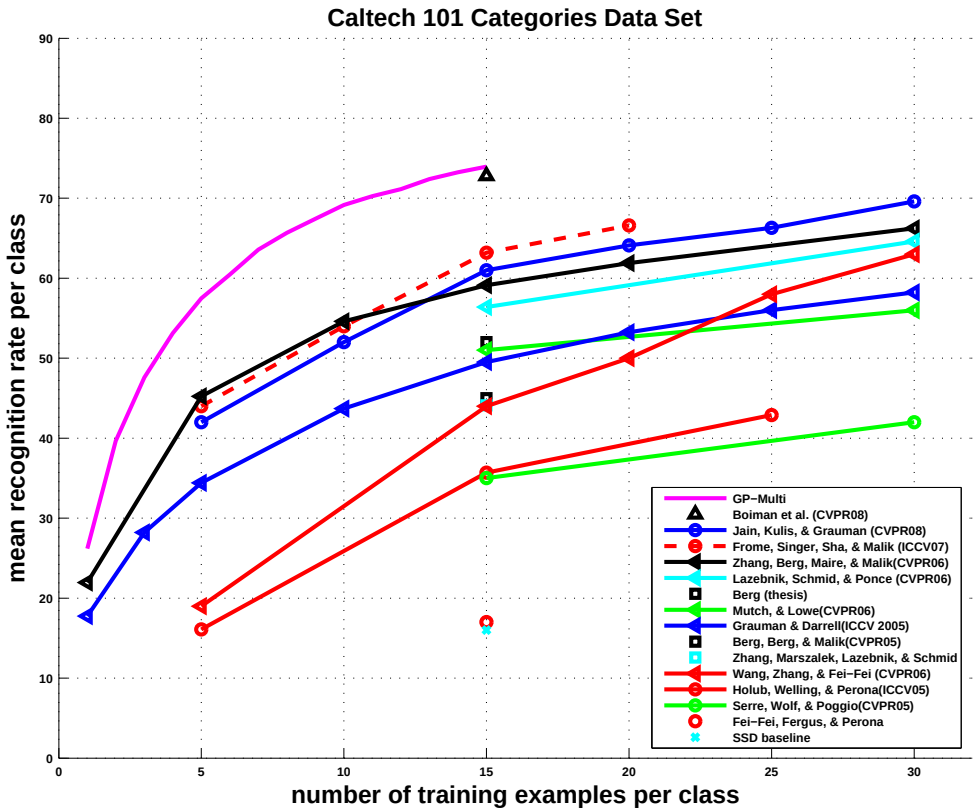


Figure: Comparison with the state of the art as in late 2008.

Other forms of MKL

Convex combination of kernels is too simple (not big boost reported), we need more complex (non-linear) combinations

- Localized comb. (the weighting varies locally) (Christioudias et al. 09)

$$\mathbf{K}^{(v)} = \mathbf{K}_{np}^{(v)} \odot \mathbf{K}_{p}^{(v)}$$

  use structure to define $\mathbf{K}_{np}^{(v)}$, e.g., low-rank

- Bayesian co-training (Yu et al. 07)

$$\mathbf{K}_c = \left[\sum_j (\mathbf{K}_j + \sigma_j^2 \mathbf{I})^{-1} \right]^{-1}$$

Other forms of MKL

Convex combination of kernels is too simple (not big boost reported), we need more complex (non-linear) combinations

- Localized comb. (the weighting varies locally) (Christioudias et al. 09)

$$\mathbf{K}^{(v)} = \mathbf{K}_{np}^{(v)} \odot \mathbf{K}_{p}^{(v)}$$

  use structure to define $\mathbf{K}_{np}^{(v)}$, e.g., low-rank

- Bayesian co-training (Yu et al. 07)

$$\mathbf{K}_c = \left[\sum_j (\mathbf{K}_j + \sigma_j^2 \mathbf{I})^{-1} \right]^{-1}$$

- Heteroscedastic Bayesian Co-training: model noise with full covariance (Christoudias et al. 09)

Check out Mario Christoudias PhD thesis for more details

Other forms of MKL

Convex combination of kernels is too simple (not big boost reported), we need more complex (non-linear) combinations

- Localized comb. (the weighting varies locally) (Christioudias et al. 09)

$$\mathbf{K}^{(v)} = \mathbf{K}_{np}^{(v)} \odot \mathbf{K}_{p}^{(v)}$$

  use structure to define $\mathbf{K}_{np}^{(v)}$, e.g., low-rank

- Bayesian co-training (Yu et al. 07)

$$\mathbf{K}_c = \left[\sum_j (\mathbf{K}_j + \sigma_j^2 \mathbf{I})^{-1} \right]^{-1}$$

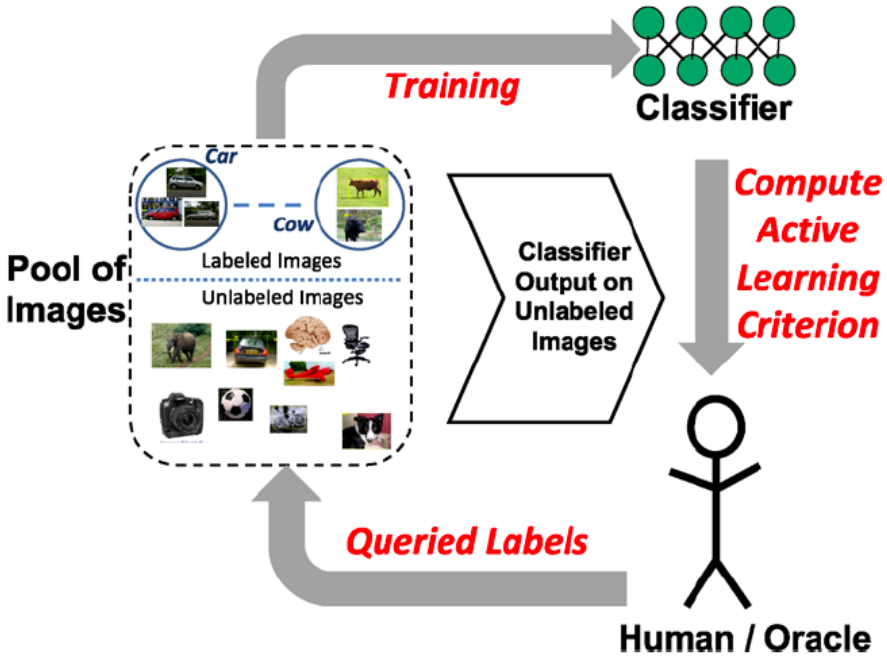
- Heteroscedastic Bayesian Co-training: model noise with full covariance (Christoudias et al. 09)

Check out Mario Christoudias PhD thesis for more details

2) Active Learning: user in the loop

- Labeling is typically an expensive process (now less with Mechanical Turk).
- Label as little as possible to reach a certain performance level.
- In **active learning**, we ask the human annotators to label not randomly, but where the classifier is more uncertain about a label.
- Trade-off between exploration and exploitation

Active Learning Algorithm



repeat

Select $\mathbf{x}_u$ to labeled using active learning criteria

As the user to label $\mathbf{x}_u$

Re-train classifier with previous labels + new label point

until budget reached

Active Learning Criteria

Let $\mathbf{X}_U$ be the pool of unlabeled data, we could use

- SVM-like criteria of distance to the margin

$$\min_{\mathbf{x}_u \in \mathbf{Y}_U} |\mathbf{y}_u^*|$$

- Most uncertain point (i.e., variance)

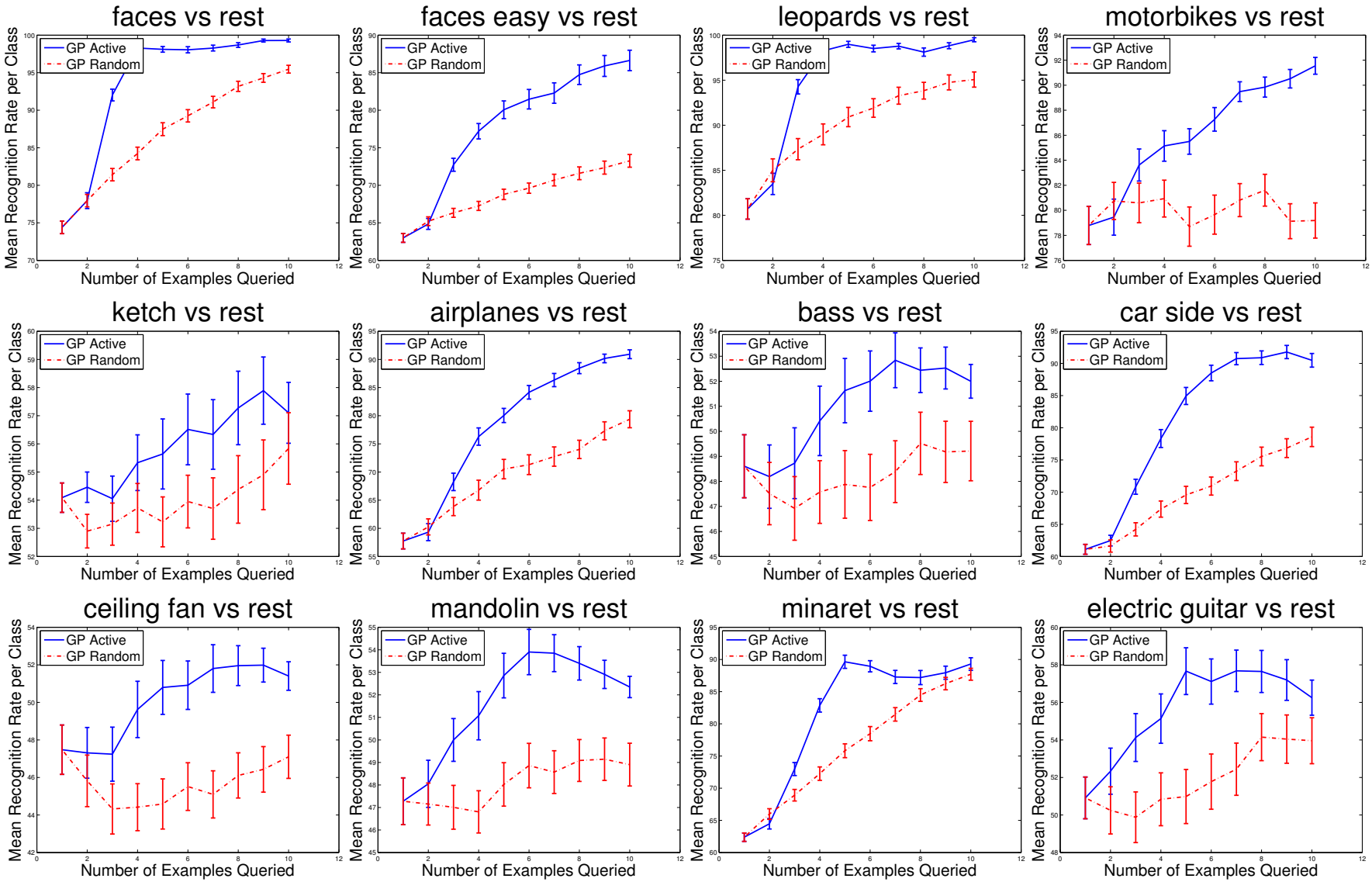
$$\max_{\mathbf{x}_u \in \mathbf{Y}_U} \Sigma_u^*$$

- Trade off between exploitation and exploration

$$\min_{\mathbf{x}_u \in \mathbf{Y}_U} \frac{|\mathbf{y}_u^*|}{\sqrt{\Sigma_u^* + \sigma^2}}$$

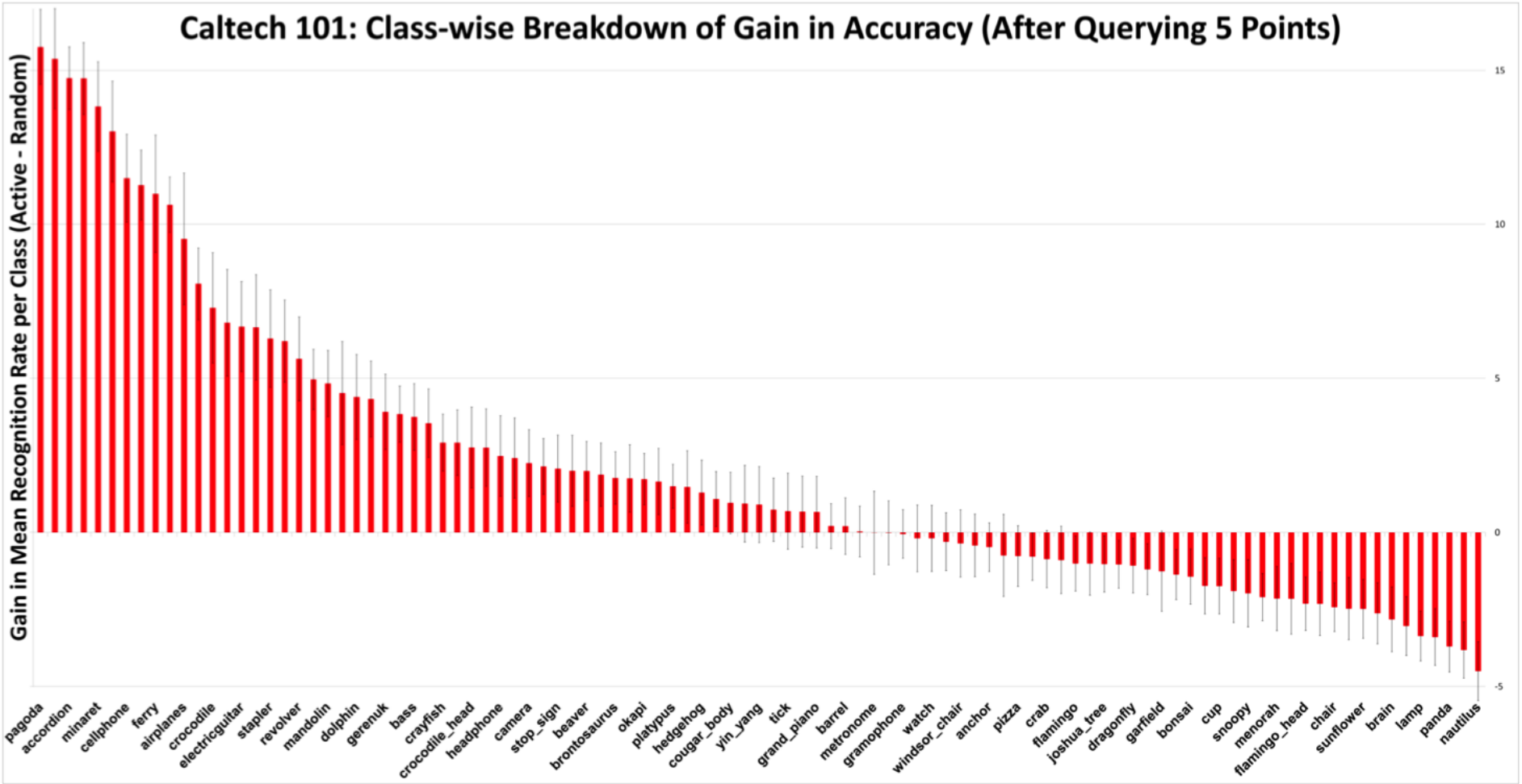
Some examples on Caltech 101

[A. Kapoor, K. Graumann, R. Urtasun and T. Darrell, ICCV 2007]



Does it always help?

[A. Kapoor, K. Graumann, R. Urtasun and T. Darrell, ICCV 2007]



3) Optimization non-differentiable functions

Supposed you have a function that you want to optimize, but it is **non-differentiable** and also **computationally expensive** to evaluate, you can

- Discretize your space and evaluate discretized values in a grid (combinatorial)
- Randomly sample your parameters
- Utilize "active learning" style analysis and GPs to query where to look

GPs as an optimization tool

[N. Srinivas, A. Krause, S. Kakade and M. Seeger, ICML 2010]

Suppose we want to compute $\max f(x)$, we can simply

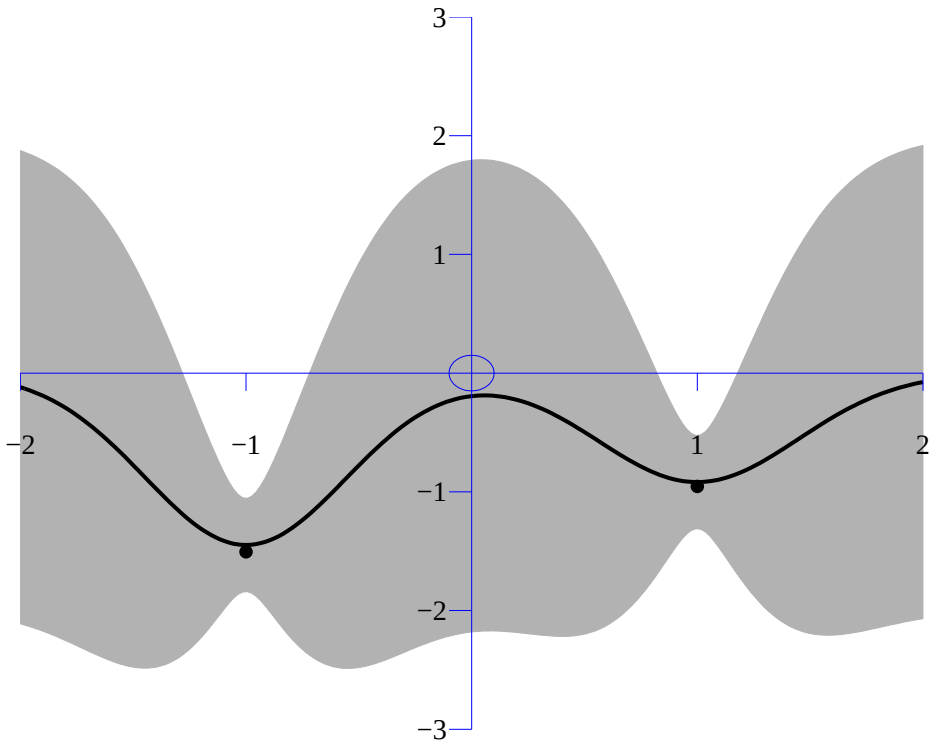
repeat

Choose $\mathbf{x}_t = arg\max_{\mathbf{x}\in D} \mu_{t-1}(\mathbf{x}) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x})$

Evaluate $\mathbf{y}_t = f(\mathbf{x}_t) + \epsilon_t$

Evaluate μ_t and σ_t

until budget reached



GPs as an optimization tool

[N. Srinivas, A. Krause, S. Kakade and M. Seeger, ICML 2010]

Suppose we want to compute $\max f(x)$, we can simply

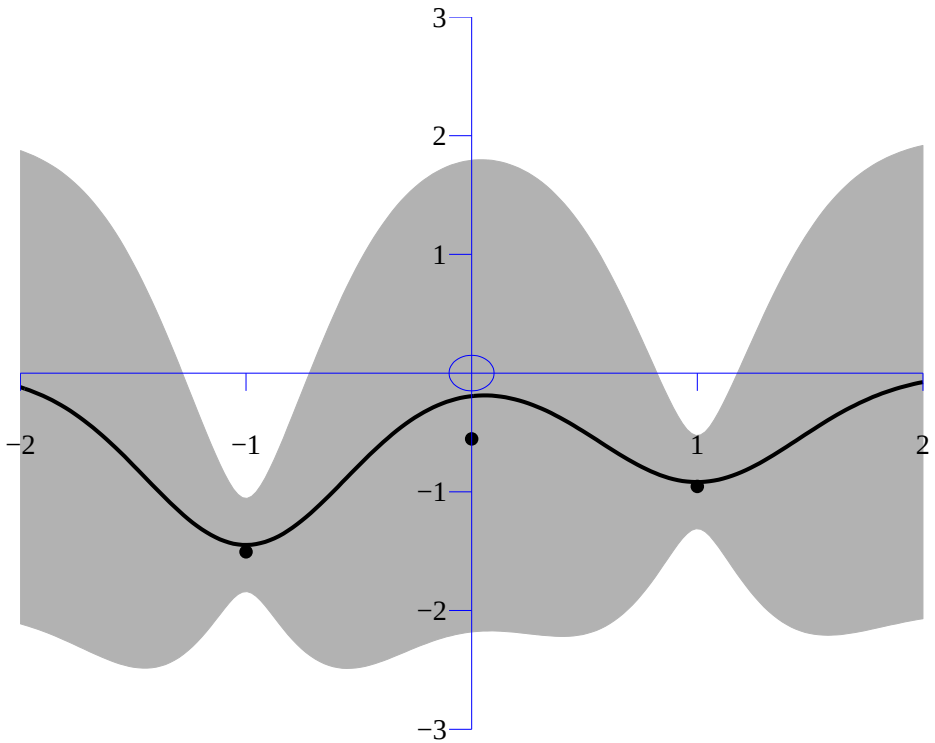
repeat

Choose $\mathbf{x}_t = arg\max_{\mathbf{x}\in D} \mu_{t-1}(\mathbf{x}) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x})$

Evaluate $\mathbf{y}_t = f(\mathbf{x}_t) + \epsilon_t$

Evaluate μ_t and σ_t

until budget reached



GPs as an optimization tool

[N. Srinivas, A. Krause, S. Kakade and M. Seeger, ICML 2010]

Suppose we want to compute $\max f(x)$, we can simply

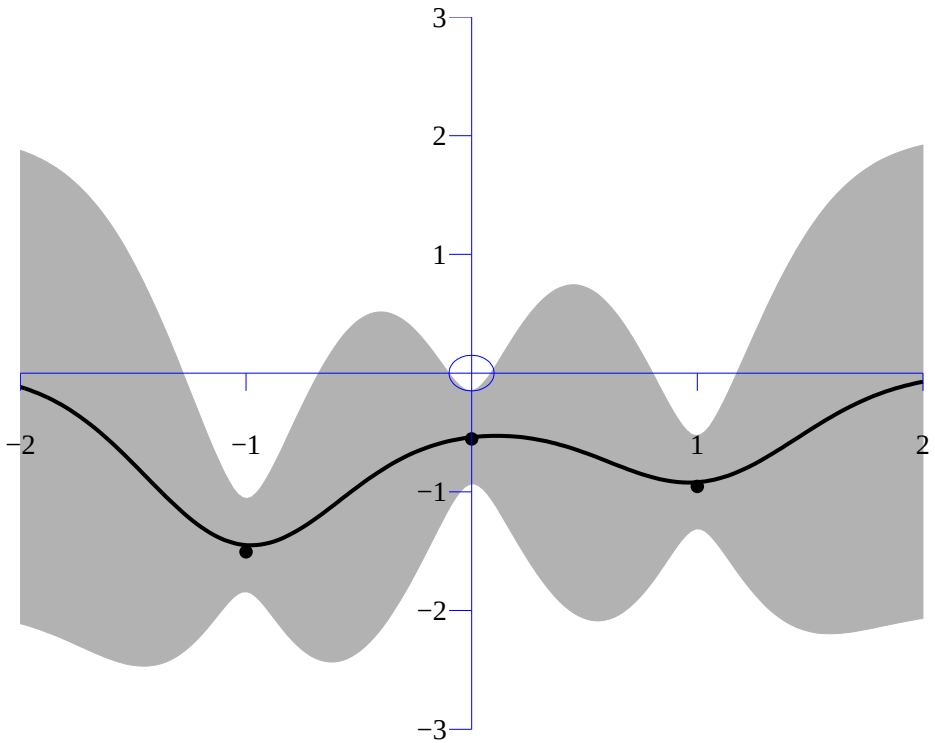
repeat

Choose $\mathbf{x}_t = arg\max_{\mathbf{x}\in D} \mu_{t-1}(\mathbf{x}) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x})$

Evaluate $\mathbf{y}_t = f(\mathbf{x}_t) + \epsilon_t$

Evaluate μ_t and σ_t

until budget reached



GPs as an optimization tool in vision

[A. Vezhnevets, V. Ferrari and J. Buhmann, CVPR 2012]

- Image segmentation in the weakly supervised setting, where the only labels are which classes are present in the scene.

$\mathbf{y} \in \{sky, building, tree\}$

- Train based on **expected agreement**, if I partition the dataset on two sets and I train on the first, it should predict the same as if I train on the second.
- This function is sum of indicator functions and thus non-differentiable.

GPs as an optimization tool in vision

[A. Vezhnevets, V. Ferrari and J. Buhmann, CVPR 2012]

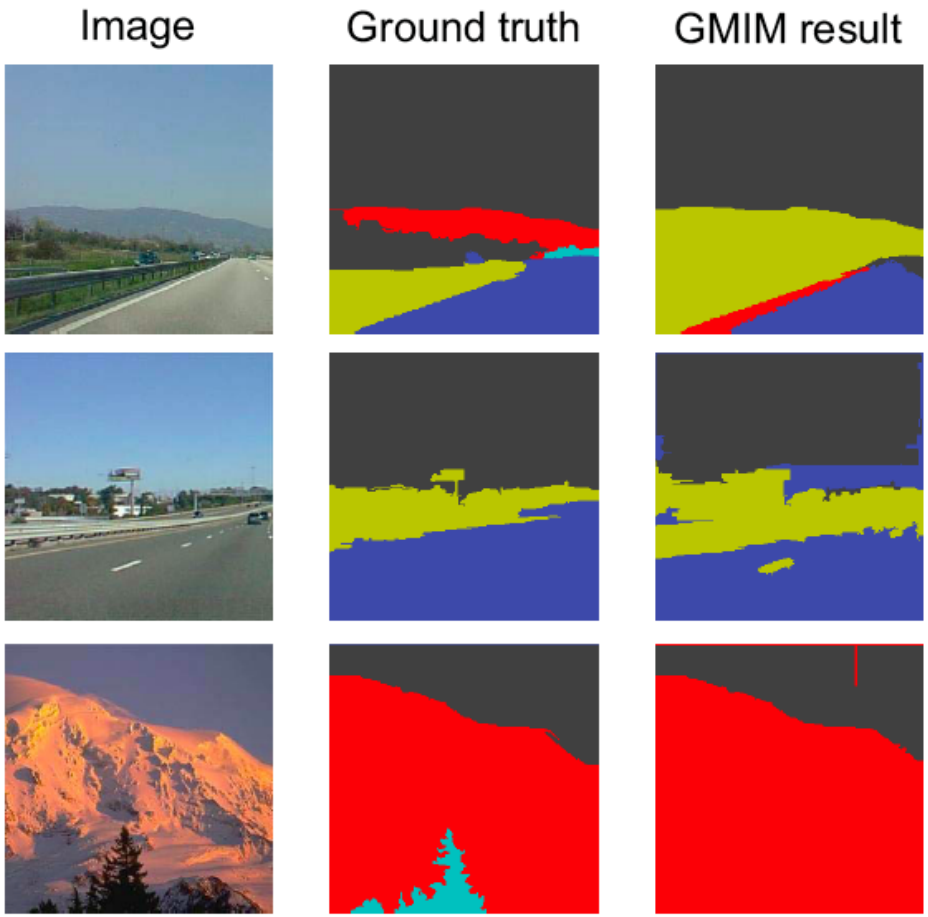
- Image segmentation in the weakly supervised setting, where the only labels are which classes are present in the scene.

$\mathbf{y} \in \{sky, building, tree\}$

- Train based on **expected agreement**, if I partition the dataset on two sets and I train on the first, it should predict the same as if I train on the second.
- This function is sum of indicator functions and thus non-differentiable.

Examples of Good Segmentations and Results

[A. Vezhnevets, V. Ferrari and J. Buhmann, CVPR 2012]



	[Tighe 10]	[Vezhnevets 11]	GMIM
supervision	fulll	weak	weak
average accuracy	29	14	21

4) Discriminative Approaches to Human Pose Estimation

- **Task:** given an image **x**, estimate the 3D location and orientation of the body parts **y**.


- We can treat this problem as a multi-output regression problem, where the input are image features, e.g., BOW, HOG, etc.
- The main challenges are
  - Poor imaging: motion blurred, occlusions, etc.
  - Need of large number of examples to represent all possible poses: represent variations in appearance and in pose.

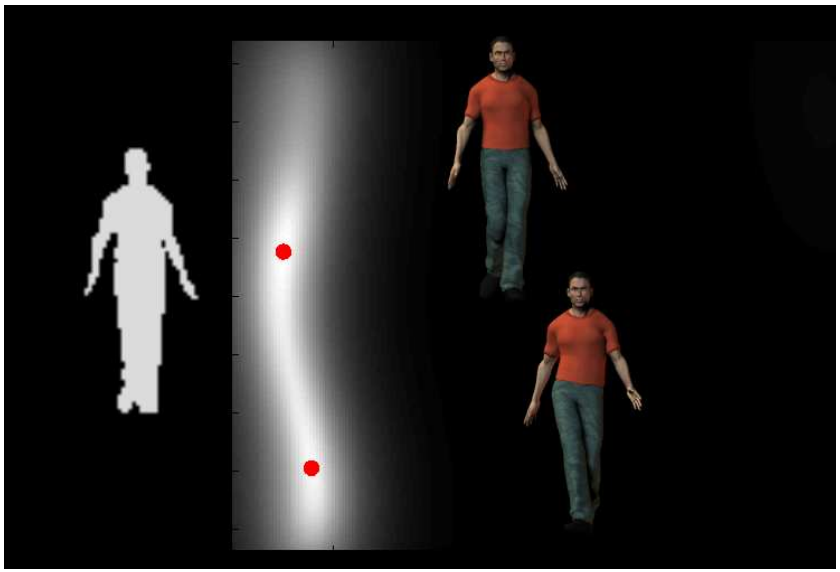
4) Discriminative Approaches to Human Pose Estimation

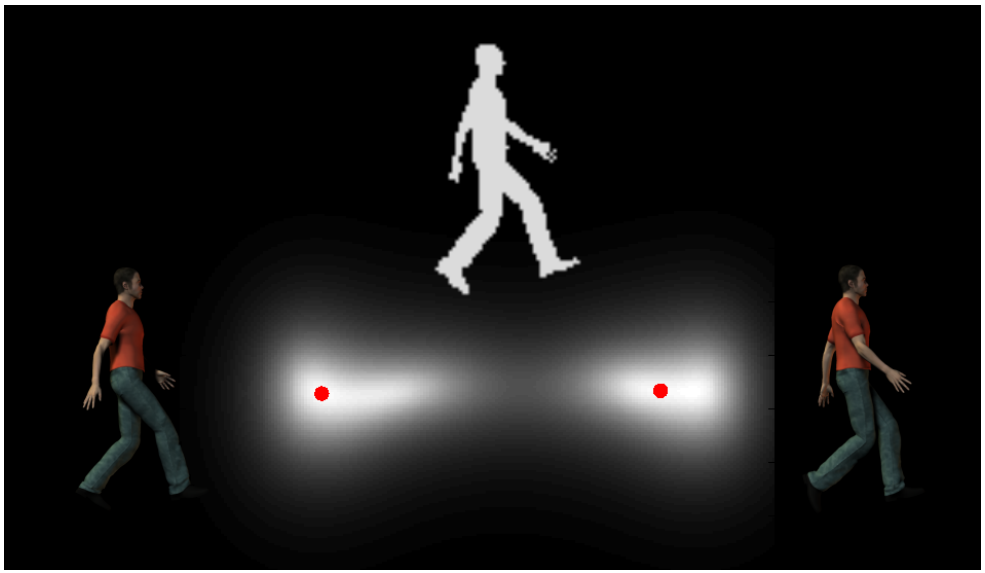
- **Task:** given an image **x**, estimate the 3D location and orientation of the body parts **y**.

- We can treat this problem as a multi-output regression problem, where the input are image features, e.g., BOW, HOG, etc.
- The main challenges are
  - Poor imaging: motion blurred, occlusions, etc.
  - Need of large number of examples to represent all possible poses: represent variations in appearance and in pose.

Challenges for GPs

- GP have complexity $\mathcal{O}(n^3)$, with n the number of examples, and cannot deal with large datasets in their standard form.
- This problem cannot be solved directly as a regression task, since the mapping is multimodal: an image observation can represent more than one pose.





Dealing with multimodal mappings

- We can represent the regression problem as a mixture of experts, where each expert is a local GP.
- The experts should be selected online to avoid the possible boundary problems of clustering.

Dealing with multimodal mappings

- We can represent the regression problem as a mixture of experts, where each expert is a local GP.
- The experts should be selected online to avoid the possible boundary problems of clustering.
- Advantages:
  - Probabilistic estimates.
  - Reliable estimation of hyperparameters
  - Strategy for pruning unnecessary examples and detecting outliers.

Dealing with multimodal mappings

- We can represent the regression problem as a mixture of experts, where each expert is a local GP.
- The experts should be selected online to avoid the possible boundary problems of clustering.
- Advantages:
  - Probabilistic estimates.
  - Reliable estimation of hyperparameters
  - Strategy for pruning unnecessary examples and detecting outliers.
- Fast solution with up to millions of examples if combined with fast NN retrieval, e.g., LSH.

Dealing with multimodal mappings

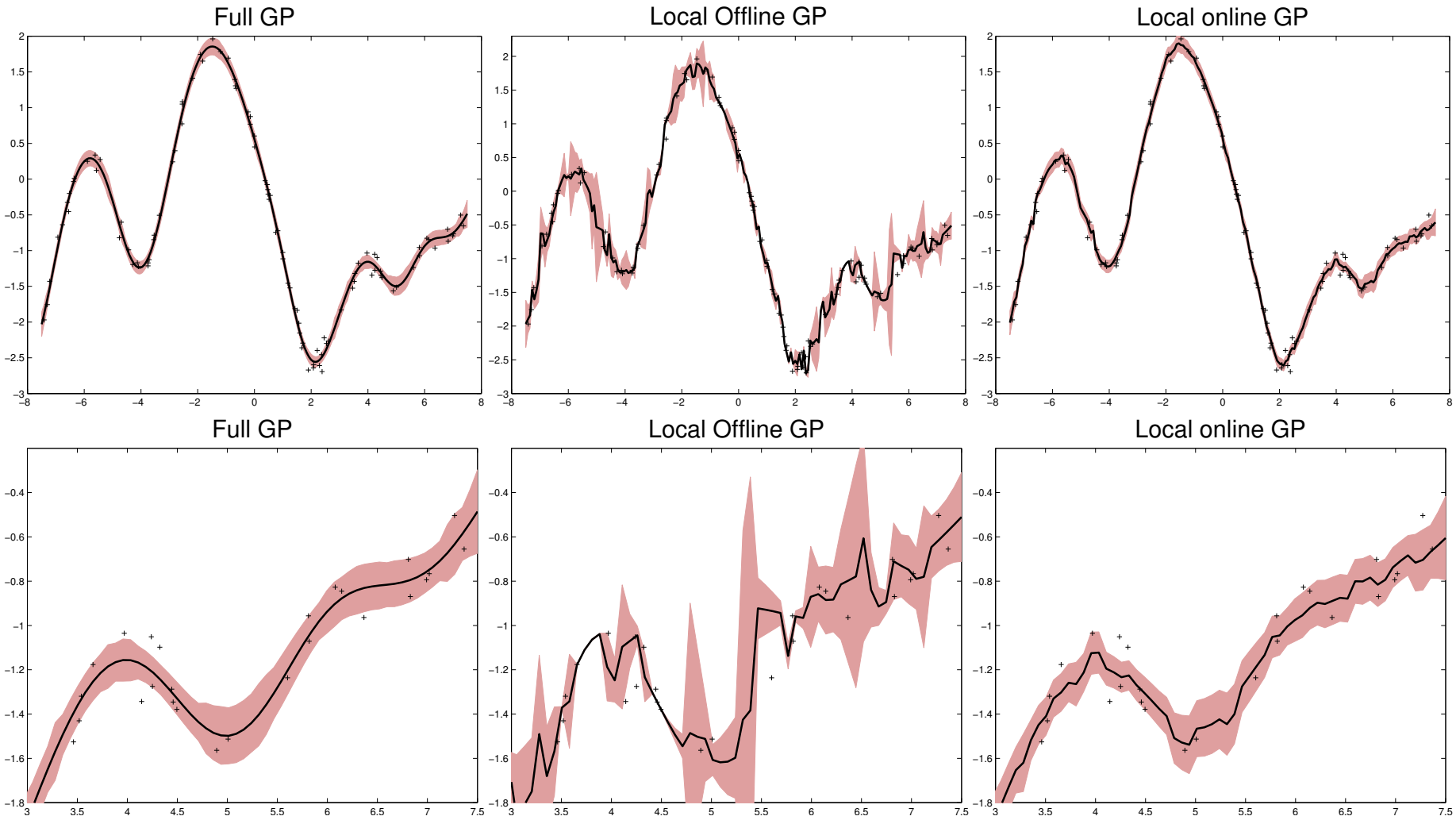
- We can represent the regression problem as a mixture of experts, where each expert is a local GP.
- The experts should be selected online to avoid the possible boundary problems of clustering.
- Advantages:
  - Probabilistic estimates.
  - Reliable estimation of hyperparameters
  - Strategy for pruning unnecessary examples and detecting outliers.
- Fast solution with up to millions of examples if combined with fast NN retrieval, e.g., LSH.

Online Algorithm

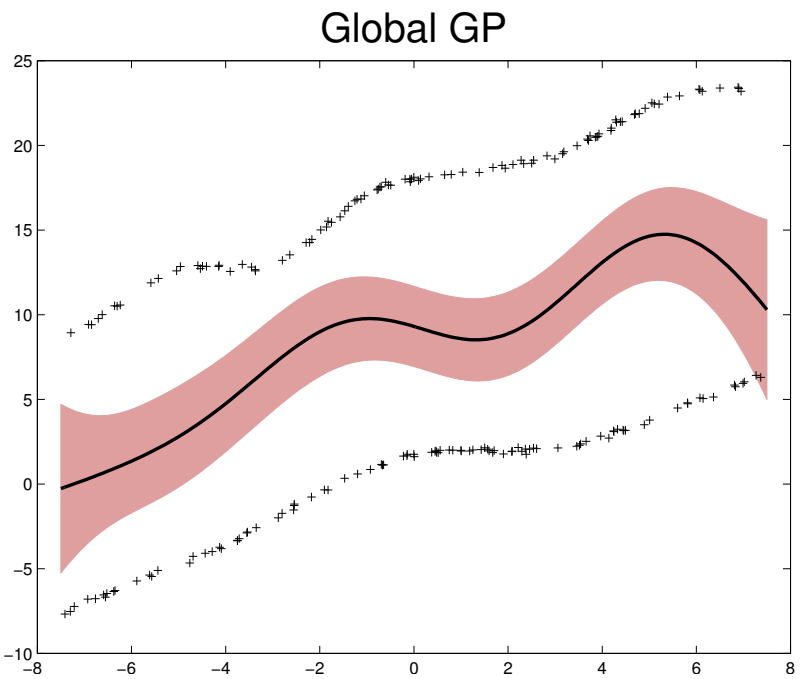
ONLINE: Inference of test point $\mathbf{x}_*$
T: number of experts, S: size of each expert

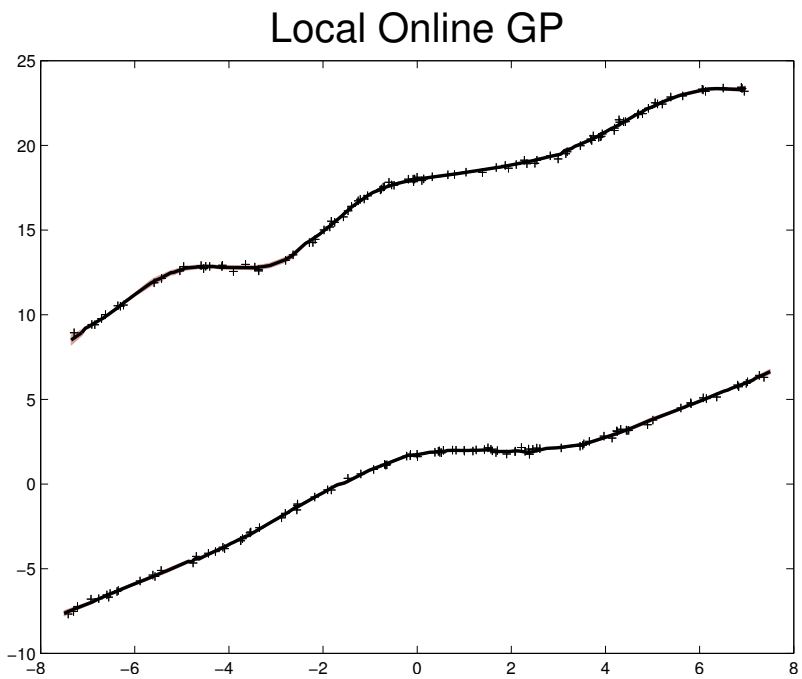
Find NN in $\mathbf{x}$ of $\mathbf{x}_*$
Find Modes in $\mathbf{y}$ of the NN retrieved
for $i = 1 \dots T$ **do**
    Create a local GP for each mode i
    Retrieve hyper-parameters
    Compute mean μ and variance σ
end for
$p(\mathbf{f}_*|\mathbf{y}) \approx \sum_{i=1}^{T} \pi_i \mathcal{N}(\mu_i, \sigma_i^2)$

Online vs Clustering



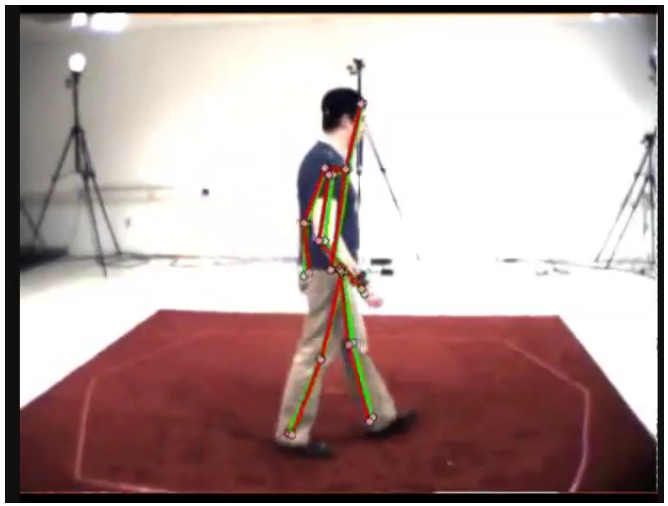
Single GP vs Mixture of Online GPs





Results: Humaneva

[R. Urtasun and T. Darrell, CVPR 2008]



	walk	jog	box	mono.	discrim.	dyn.
Lee et al. I	3.4	–	–	yes	no	no
Lee et al. II	**3.1**	–	–	yes	no	yes
Pope	4.53	4.38	9.43	yes	yes	no
Muendermann et al.	5.31	–	4.54	no	no	yes
Li et al.	–	–	20.0	yes	no	yes
Brubaker et al.	10.4	–	–	yes	no	yes
Our approach	3.27	**3.12**	**3.85**	yes	yes	no

Table: Comparison with state of the art (error in cm).

- Caviat: Oracle has to select the optimal mixture component

5) Flow Estimation with Gaussian Process

- Model a trajectory as a continuous dense flow field from a sparse set of vector sequences using Gaussian Process Regression
- Each velocity component modeled with an independent GP
- The flow can be expressed as

$$\phi(\mathbf{x}) = \mathbf{y}^{(u)}(\mathbf{x})\mathbf{i} + \mathbf{y}^{(v)}(\mathbf{x})\mathbf{j} + \mathbf{y}^{(t)}(\mathbf{x})\mathbf{k} \in \Re^3$$

  where $\mathbf{x} = (u, v, t)$

- Difficulties:
  - How to model a GPRF from different trajectories, which may have different lengths
  - How to handle multiple GPRF models trained from different numbers of trajectories with heterogeneous scales and frame rates

5) Flow Estimation with Gaussian Process

- Model a trajectory as a continuous dense flow field from a sparse set of vector sequences using Gaussian Process Regression
- Each velocity component modeled with an independent GP
- The flow can be expressed as

$$\phi(\mathbf{x}) = \mathbf{y}^{(u)}(\mathbf{x})\mathbf{i} + \mathbf{y}^{(v)}(\mathbf{x})\mathbf{j} + \mathbf{y}^{(t)}(\mathbf{x})\mathbf{k} \in \Re^3$$

  where $\mathbf{x} = (u, v, t)$
- Difficulties:
  - How to model a GPRF from different trajectories, which may have different lengths
  - How to handle multiple GPRF models trained from different numbers of trajectories with heterogeneous scales and frame rates
- Solution: normalize the length of the tracks before modeling with a GP, as well as the number of samples
- Classification based on the likelihood for each class

5) Flow Estimation with Gaussian Process

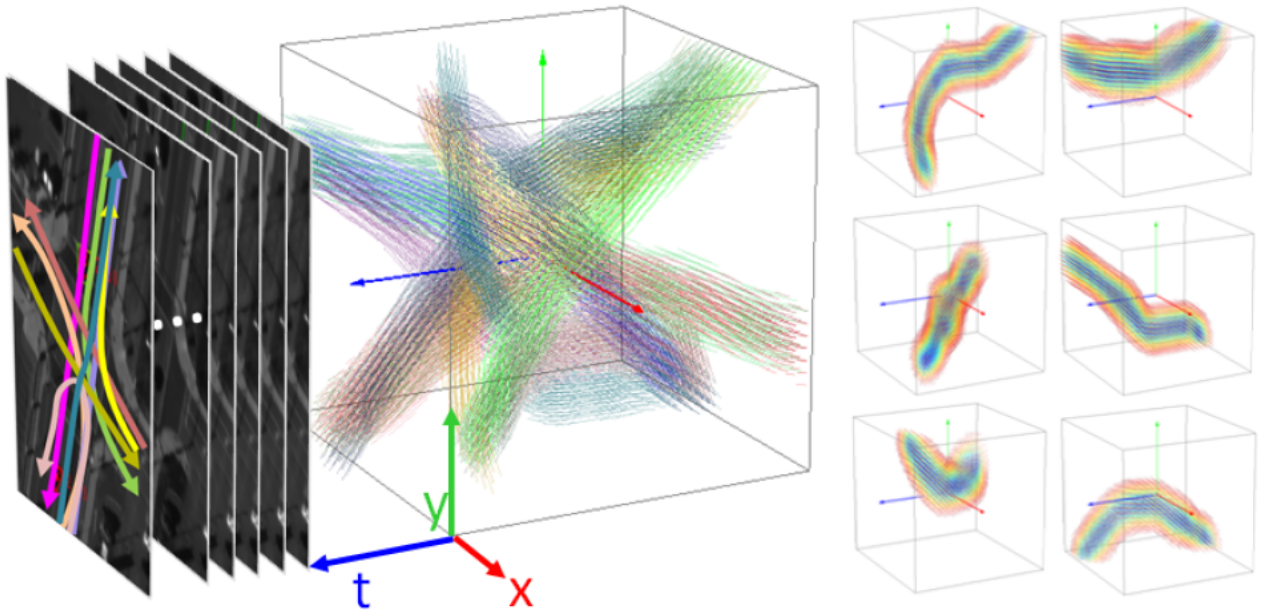
- Model a trajectory as a continuous dense flow field from a sparse set of vector sequences using Gaussian Process Regression
- Each velocity component modeled with an independent GP
- The flow can be expressed as

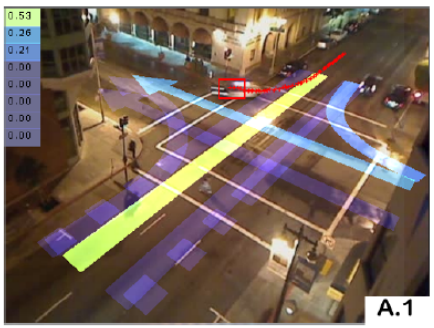
$$\phi(\mathbf{x}) = \mathbf{y}^{(u)}(\mathbf{x})\mathbf{i} + \mathbf{y}^{(v)}(\mathbf{x})\mathbf{j} + \mathbf{y}^{(t)}(\mathbf{x})\mathbf{k} \in \Re^3$$

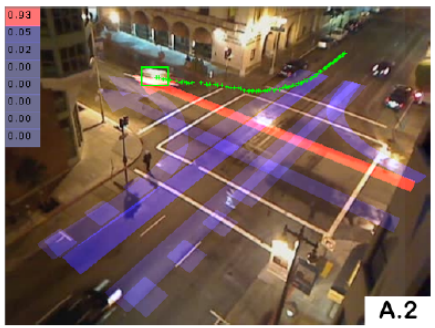
  where $\mathbf{x} = (u, v, t)$
- Difficulties:
  - How to model a GPRF from different trajectories, which may have different lengths
  - How to handle multiple GPRF models trained from different numbers of trajectories with heterogeneous scales and frame rates
- Solution: normalize the length of the tracks before modeling with a GP, as well as the number of samples
- Classification based on the likelihood for each class

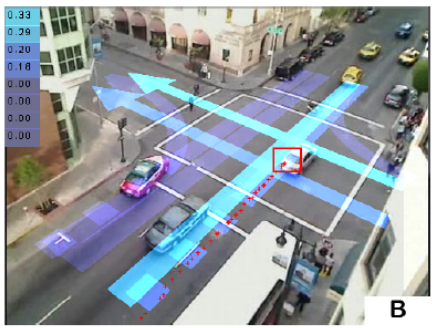
Flow Classification and Anomaly Detection

[K. Kim, D. Lee and I. Essa, ICCV 2011]









Detecting Regions of Interest

- Detect the regions of interests in moving camera views of dynamic scenes with multiple moving objects
- Important for cheap streaming of events on the internet, e.g., NBA playoffs on TNT
- Extract a global motion tendency that reflects the scene context by tracking movements of objects in the scene

Detecting Regions of Interest

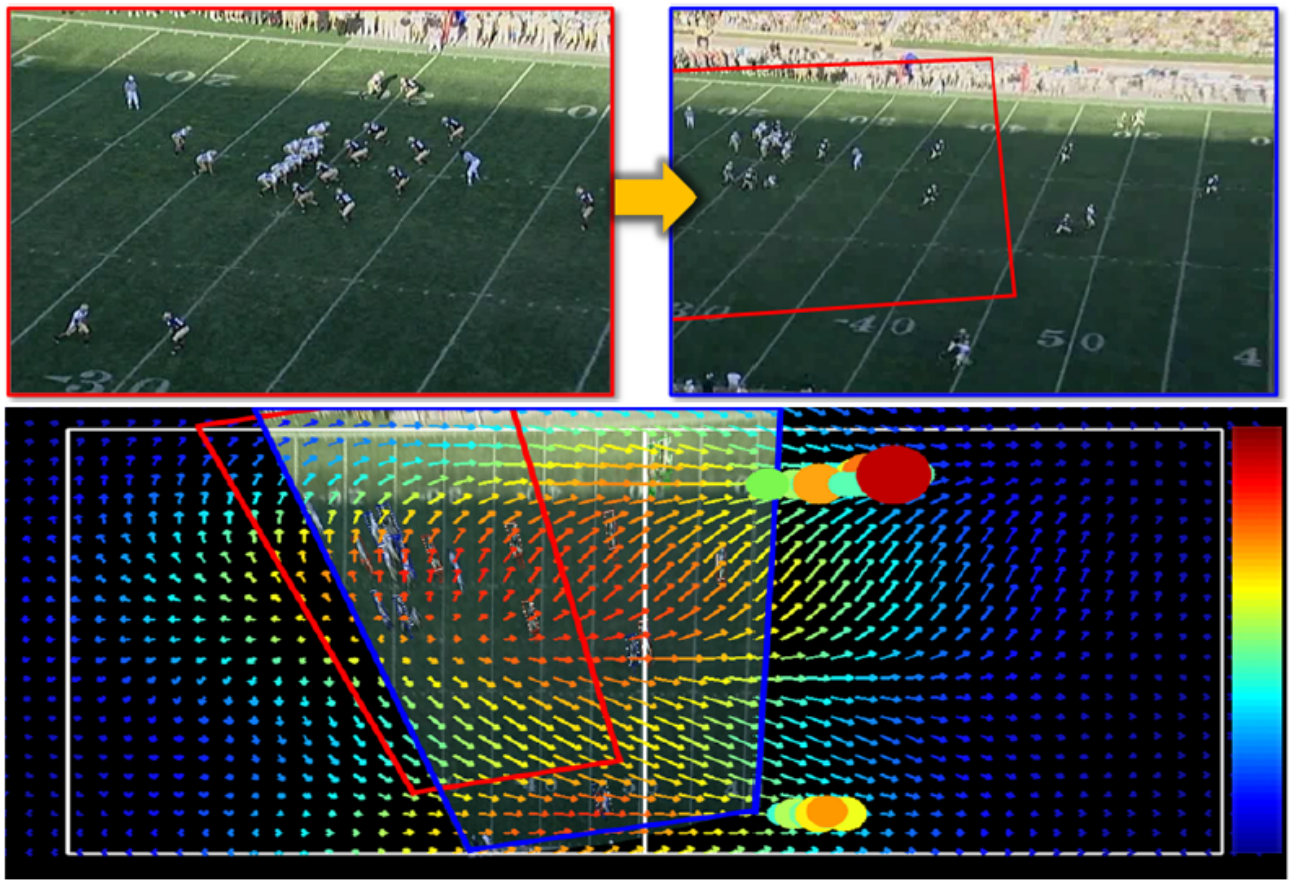
- Detect the regions of interests in moving camera views of dynamic scenes with multiple moving objects
- Important for cheap streaming of events on the internet, e.g., NBA playoffs on TNT
- Extract a global motion tendency that reflects the scene context by tracking movements of objects in the scene
- Use GPs to represent the extracted motion tendency as a stochastic vector field.
- Use the stochastic field for predicting important future regions of interest as the scene evolves dynamically

Detecting Regions of Interest

- Detect the regions of interests in moving camera views of dynamic scenes with multiple moving objects
- Important for cheap streaming of events on the internet, e.g., NBA playoffs on TNT
- Extract a global motion tendency that reflects the scene context by tracking movements of objects in the scene
- Use GPs to represent the extracted motion tendency as a stochastic vector field.
- Use the stochastic field for predicting important future regions of interest as the scene evolves dynamically

Flow Classification and Anomaly Detection

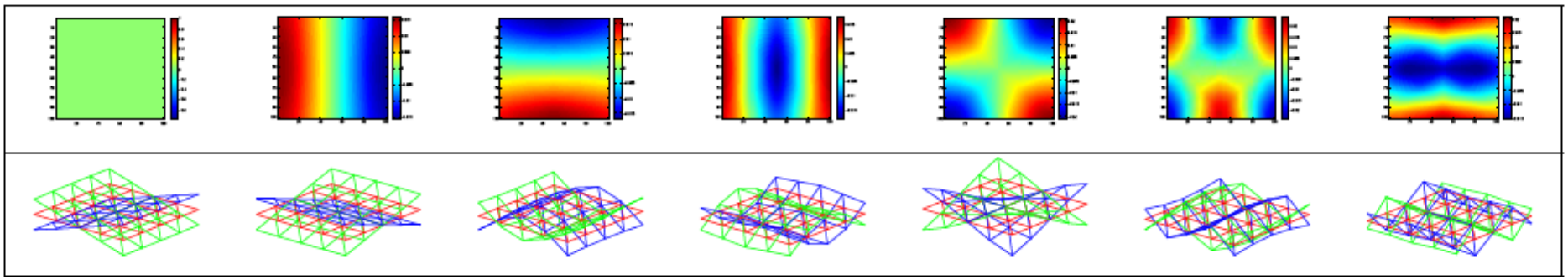
[K. Kim, D. Lee and I. Essa, CVPR 2012]



6) Shape from Shading

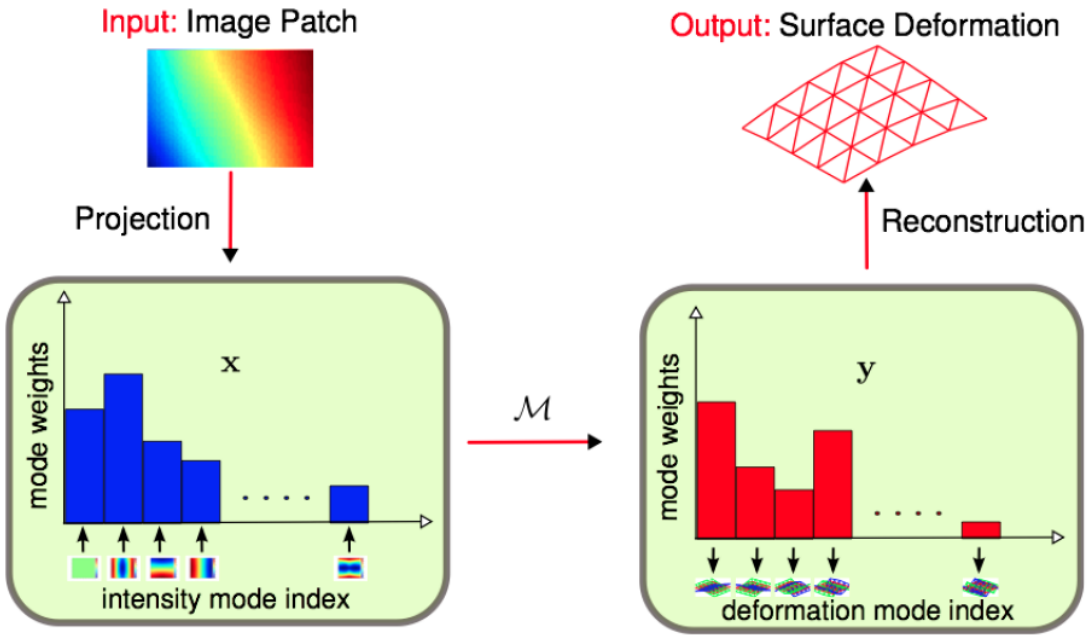
- 3D shapes were obtained using a motion capture system and taking small 5×5 patches from the reconstructed shapes
- Corresponding intensities were obtained using a Computer Graphics software and the calibrated lighting environment
- To reduce the dimensionality, performed PCA on the 3D patch shapes and on the patch intensities separately.
- Local intensities and shapes are then represented as

$$\mathbf{I} = \mathbf{I}_0 + \sum_{i=1}^{N_I} x_i \mathbf{I}_i \ , \qquad \mathbf{D} = \mathbf{D}_0 + \sum_{i=1}^{N_D} y_i \mathbf{D}_i$$



Gaussian Process Mapping

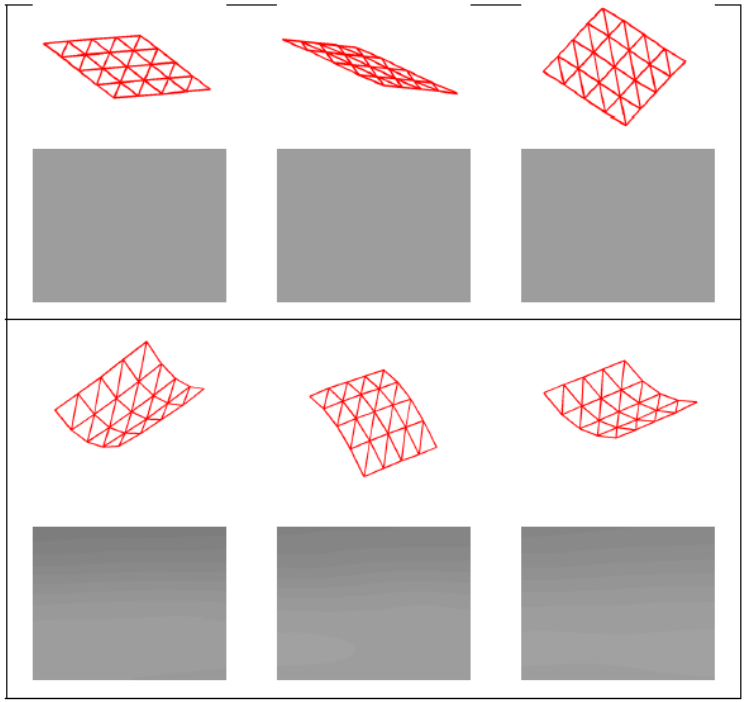
- We want to learn a mapping from intensity to 3D patch shape
- This is equivalent to learn a mapping from **x** to **y**



- Use GPs to learn this mapping
- Illumination modeled as weighted sum of spherical harmonics
- Illumination parameters estimated using a light probe

Handling Multimodality

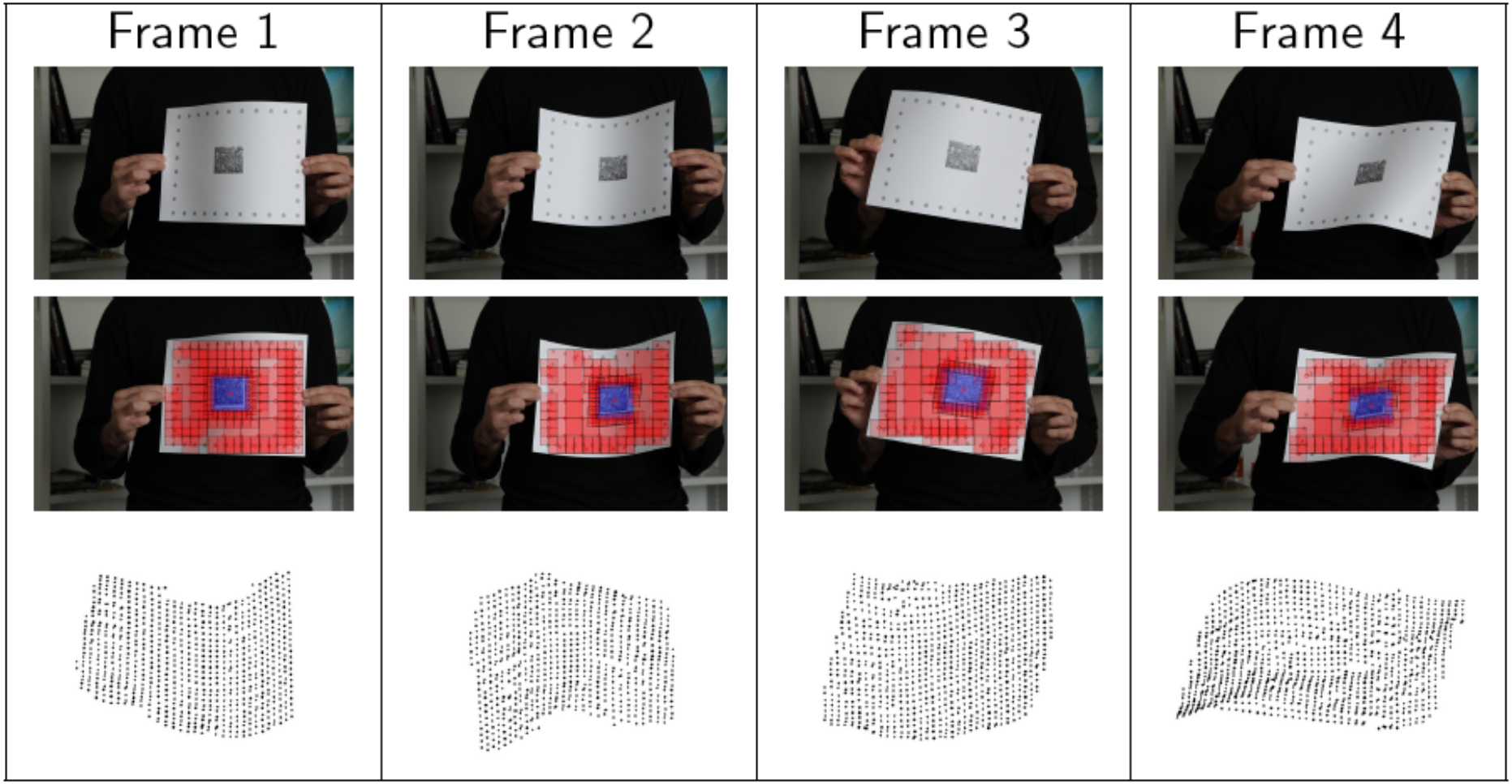
- Unfortunately, this mapping is multimodal (i.e., similar intensities may correspond to different 3D shapes)



- These ambiguities are strongly related to the first two shape component, which define out-of-plane rotation
- Use local GPs (Urtasun et al., 08) to handle multiple modes
- Utilize an MRF to patch the local patches.

Real Data Experiments

[A. Varol, A. Shaji, M. Salzmann and P. Fua, PAMI 2011]



7) 3D Shape Recovery for Online Shopping

- Interactive system for quickly modelling 3D body shapes from a single image
- Obtain their 3D body shapes so as to try on virtual garments online

7) 3D Shape Recovery for Online Shopping

- Interactive system for quickly modelling 3D body shapes from a single image
- Obtain their 3D body shapes so as to try on virtual garments online
- Interface for users to conveniently extract anthropometric measurements from a single photo, while using readily available scene cues for automatic image rectification

7) 3D Shape Recovery for Online Shopping

- Interactive system for quickly modelling 3D body shapes from a single image
- Obtain their 3D body shapes so as to try on virtual garments online
- Interface for users to conveniently extract anthropometric measurements from a single photo, while using readily available scene cues for automatic image rectification
- GPs to predict the body parameters from input measurements while correcting the aspect ratio ambiguity resulting from photo rectification

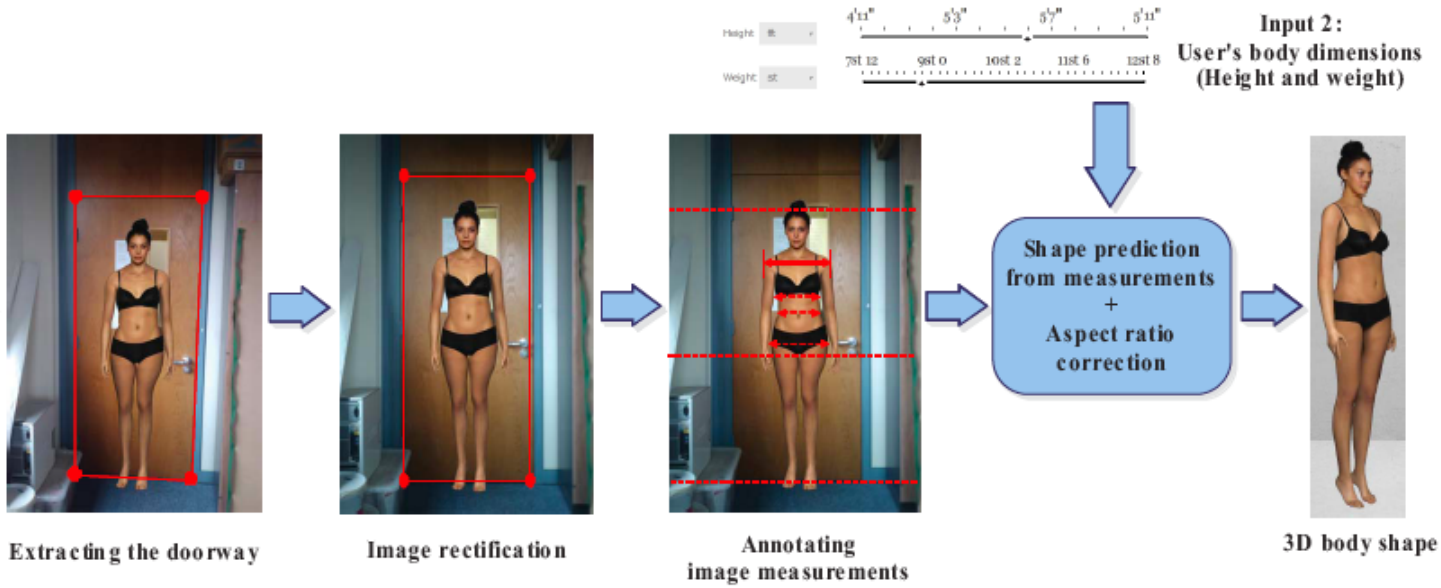
7) 3D Shape Recovery for Online Shopping

- Interactive system for quickly modelling 3D body shapes from a single image
- Obtain their 3D body shapes so as to try on virtual garments online
- Interface for users to conveniently extract anthropometric measurements from a single photo, while using readily available scene cues for automatic image rectification
- GPs to predict the body parameters from input measurements while correcting the aspect ratio ambiguity resulting from photo rectification

Creating the 3D Shape from Single Images

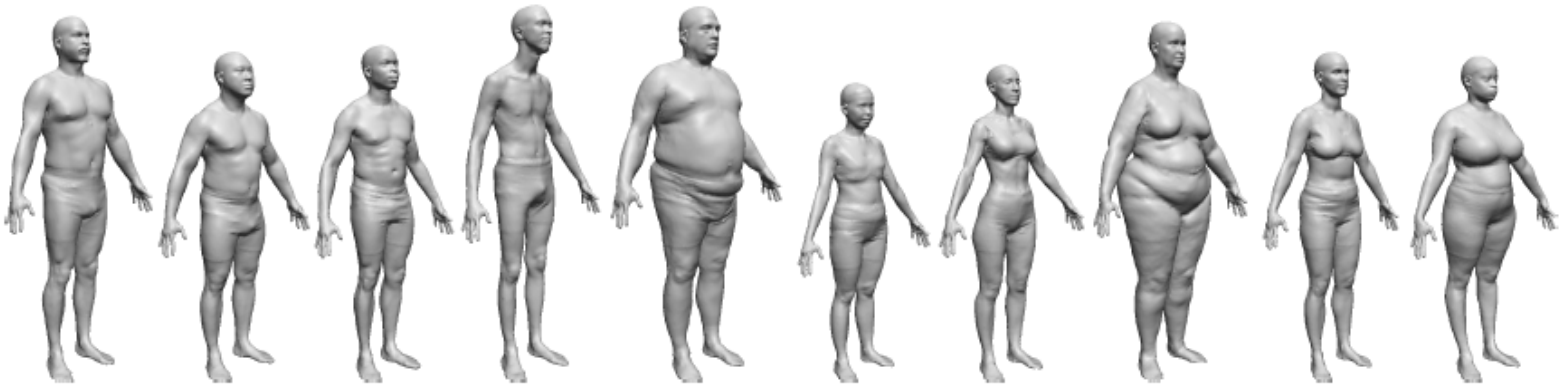
Manually annotate a set of five 2D measurements

- Well-defined by the anthropometric positions, easy to discern and unambiguous to users.
- Good correlations with the corresponding tape measurements and convey enough information for estimating the 3D body shape
- User's effort for annotation should be minimised



The role of the GPs

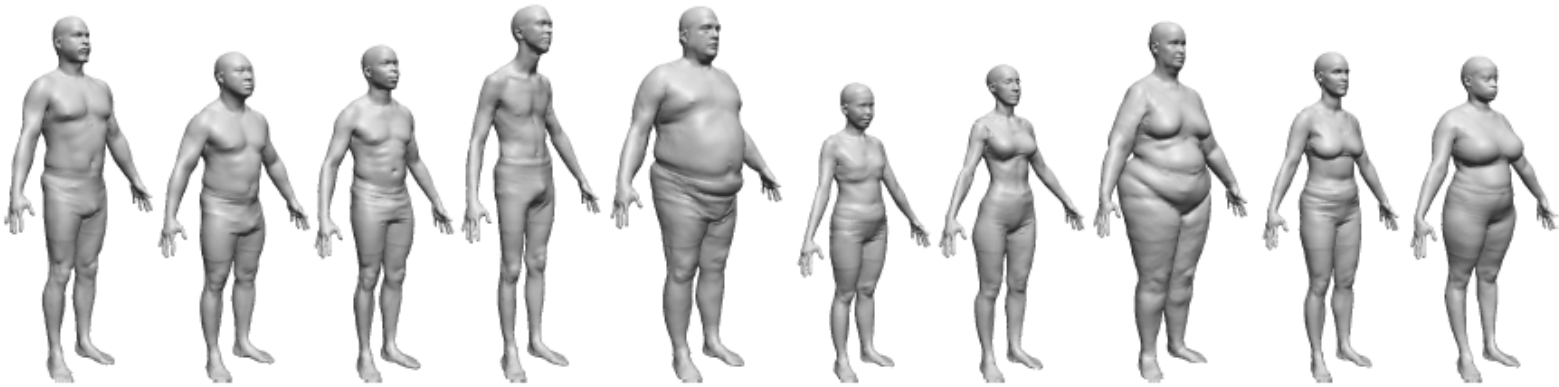
- A body shape estimator is learned to predict the 3D body shape from user's input, including both image measurements and actual measurements.
- Training set is (CAESAR) dataset (Robinette et al. 99), with 2000 bodies.



- Register each 3D instance in the dataset with a 3D morphable human body
- A 3D body is decomposed into a linear combination of body morphs

The role of the GPs

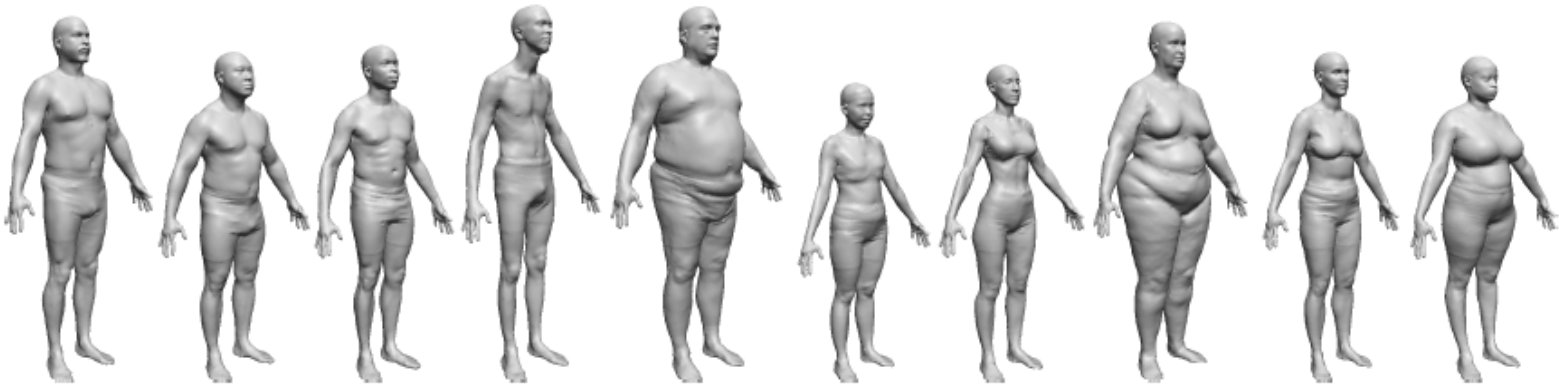
- A body shape estimator is learned to predict the 3D body shape from user's input, including both image measurements and actual measurements.
- Training set is (CAESAR) dataset (Robinette et al. 99), with 2000 bodies.



- Register each 3D instance in the dataset with a 3D morphable human body
- A 3D body is decomposed into a linear combination of body morphs
- Shape-from-measurements estimator can be formulated into a regression problem, $\mathbf{y}$ is the morph parameters and $\mathbf{x}$ is the user specified parameters.
- Multi-output done as independent predictors, each with a GP

The role of the GPs

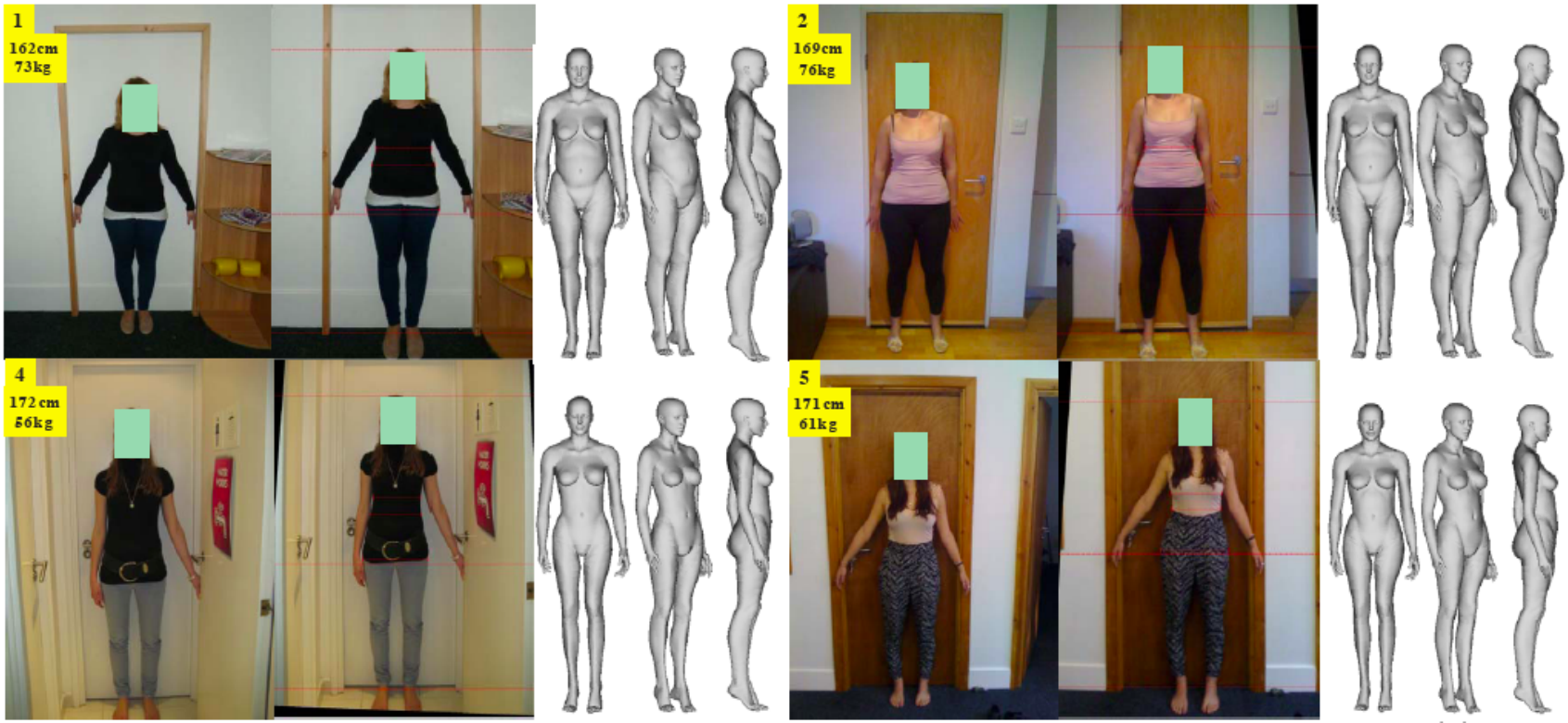
- A body shape estimator is learned to predict the 3D body shape from user's input, including both image measurements and actual measurements.
- Training set is (CAESAR) dataset (Robinette et al. 99), with 2000 bodies.



- Register each 3D instance in the dataset with a 3D morphable human body
- A 3D body is decomposed into a linear combination of body morphs
- Shape-from-measurements estimator can be formulated into a regression problem, $\mathbf{y}$ is the morph parameters and $\mathbf{x}$ is the user specified parameters.
- Multi-output done as independent predictors, each with a GP

Online Shopping

[Y. Chen and D. Robertson and R. Cipolla, BMVC 2011]



	Chest	Waist	Hips	Inner leg length
Error(cm)	1.52 ± 1.36	1.88 ± 1.06	3.10 ± 1.86	0.79 ± 0.90