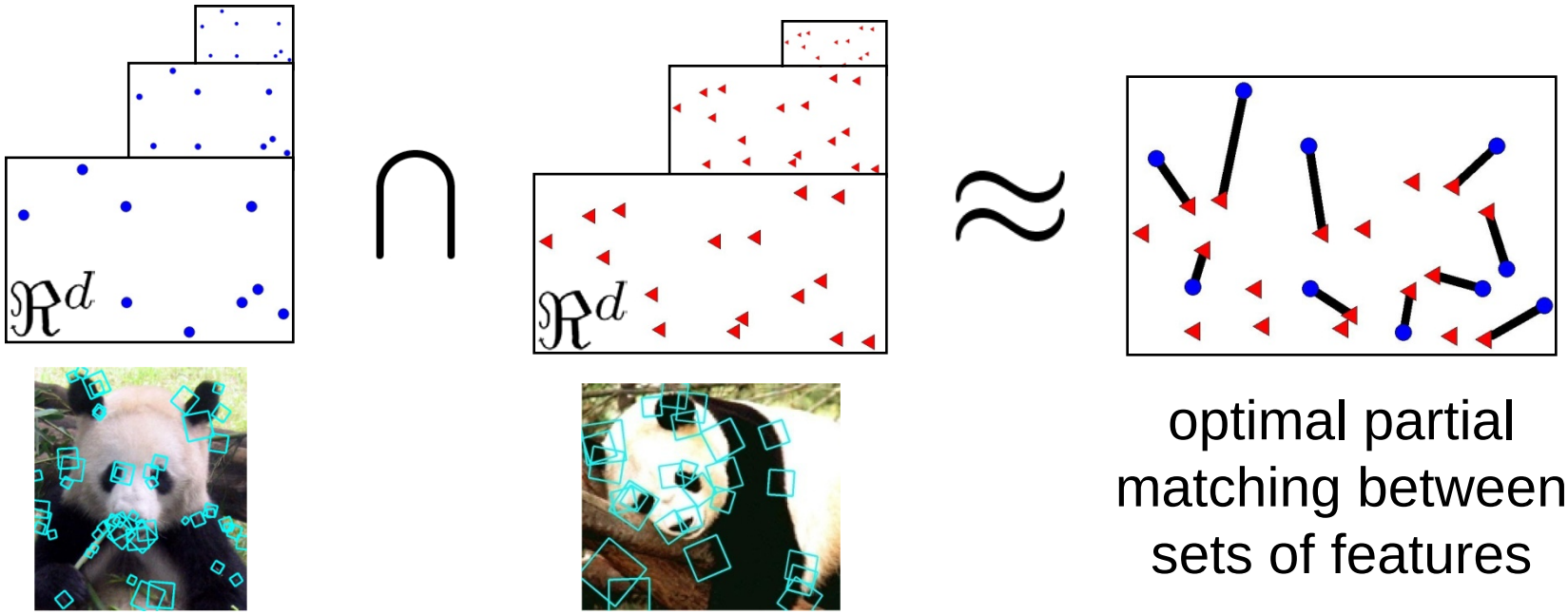
Advanced Representation and Learning

Vinay P. Namboodiri

- Slide Credit to James Hays, Hakan Bilen, Kai Yu and Derek Hoiem

Last class: Pyramid match kernel



optimal partial matching between sets of features

$$K_\Delta\left(\Psi(\mathbf{X}), \Psi(\mathbf{Y})\right) = \sum_{i=0}^{L} \underbrace{\frac{1}{2^i}}_{\text{difficulty of a match at level i}} \underbrace{\left(\mathcal{I}\left(H_i(\mathbf{X}), H_i(\mathbf{Y})\right) - \mathcal{I}\left(H_{i-1}(\mathbf{X}), H_{i-1}(\mathbf{Y})\right)\right)}_{\text{number of new matches at level i}}$$

Last Class: Spatial Pyramid

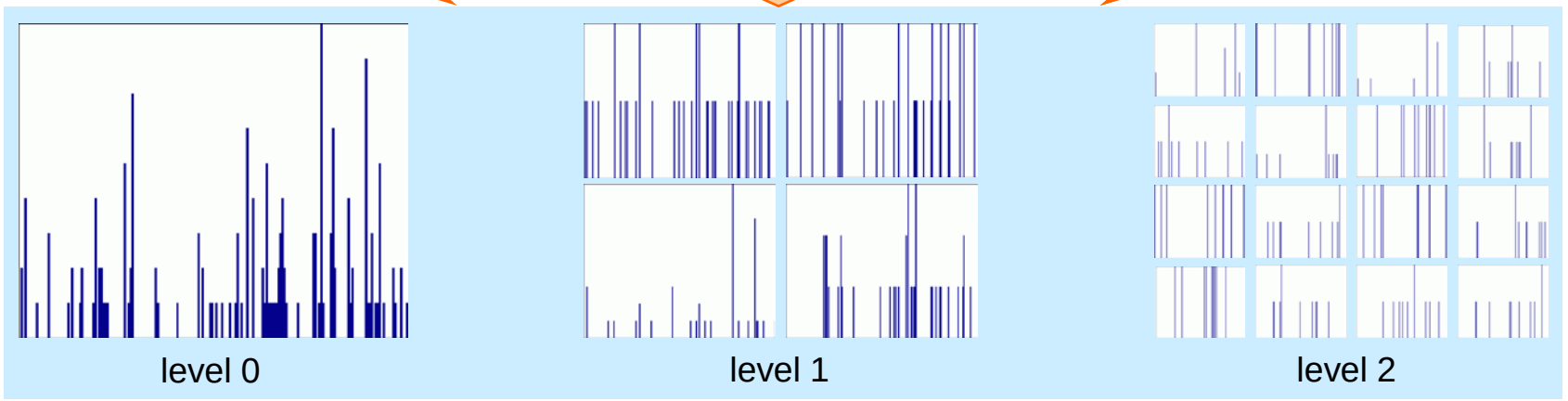


Image Categorization

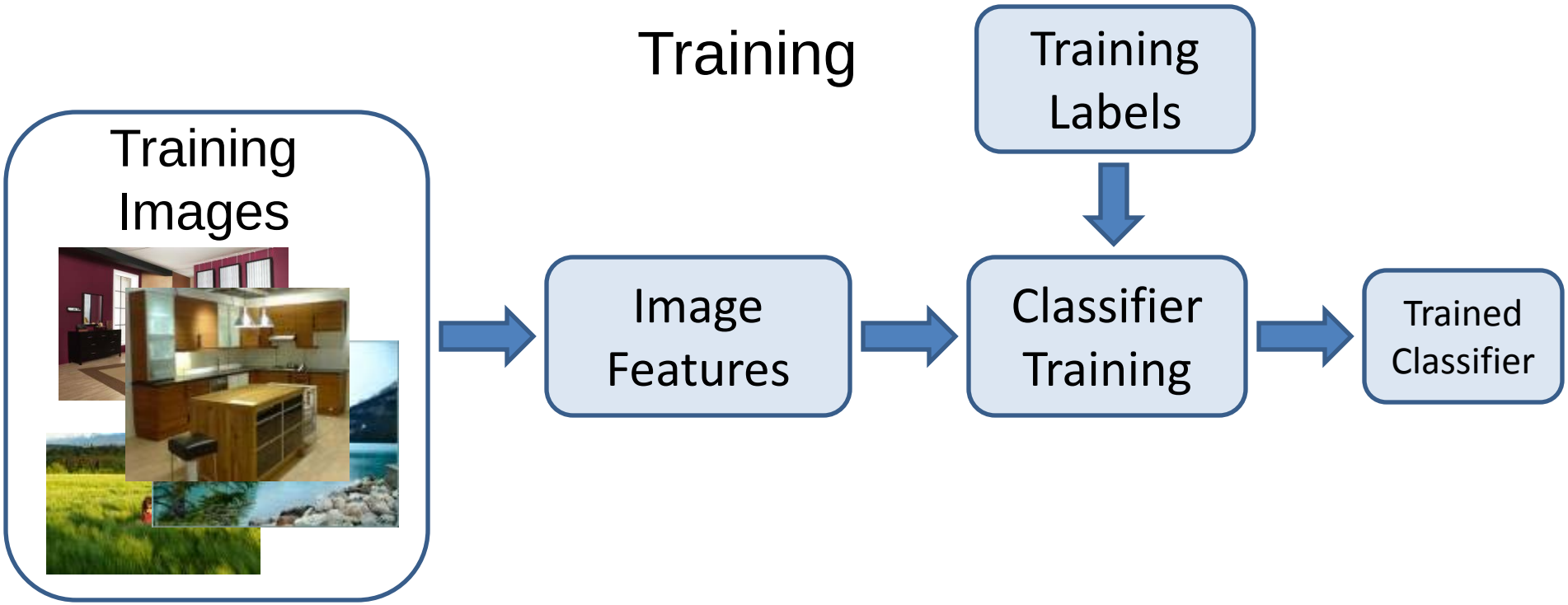


Image Categorization

Training

Training Images

Training Labels

Image Features

Classifier Training

Trained Classifier

Testing

Test Image

Image Features

Trained Classifier

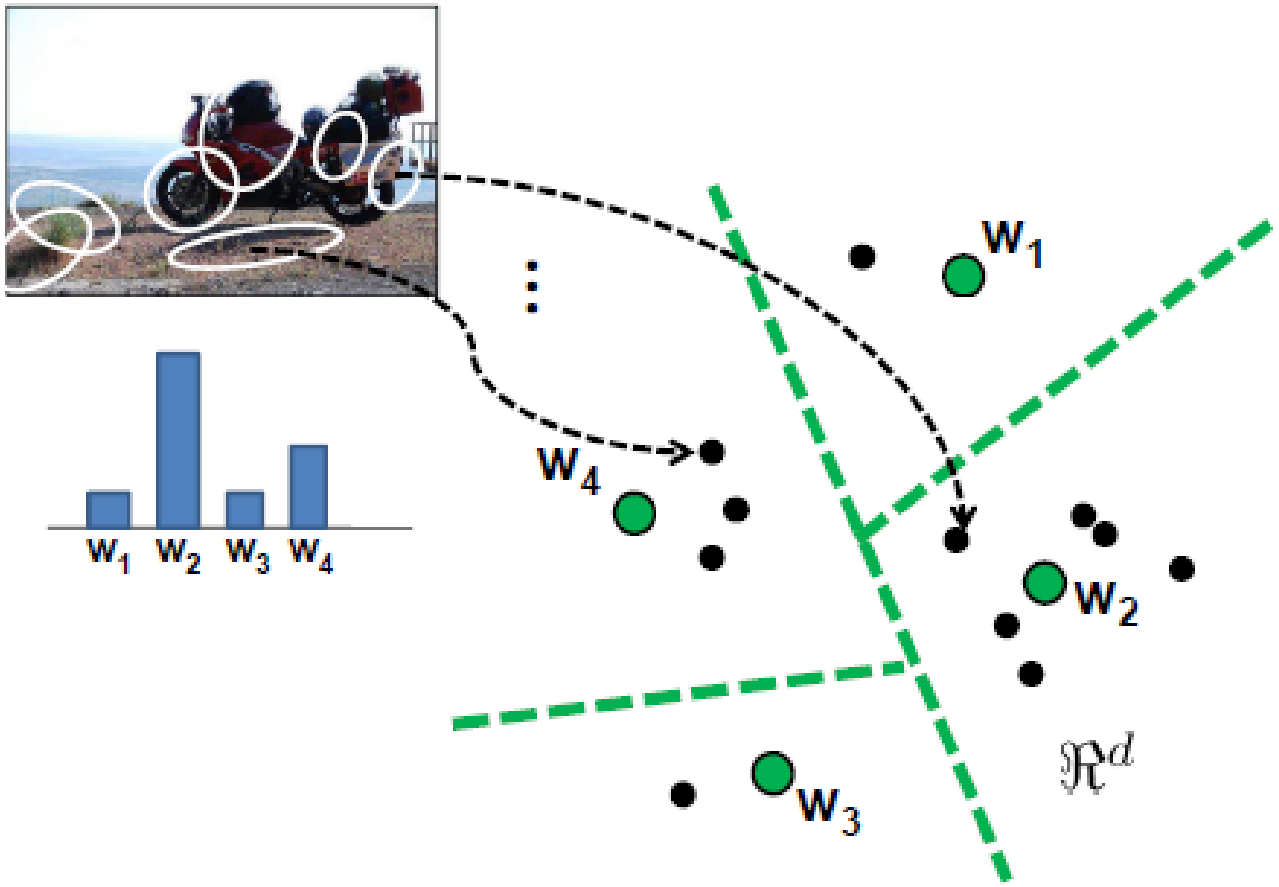
Prediction

Outdoor

Why do good recognition systems go bad?

- E.g. Why isn't our Bag of Words classifier at 90% instead of 70%?
- Training Data
  - Huge issue, but not necessarily a variable you can manipulate.
- Learning method
  - Probably not such a big issue, unless you're learning the representation (e.g. deep learning).
- Representation
  - Are the local features themselves lossy
  - What about feature quantization? That's VERY lossy.

Standard Kmeans Bag of Words



http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Today's Class

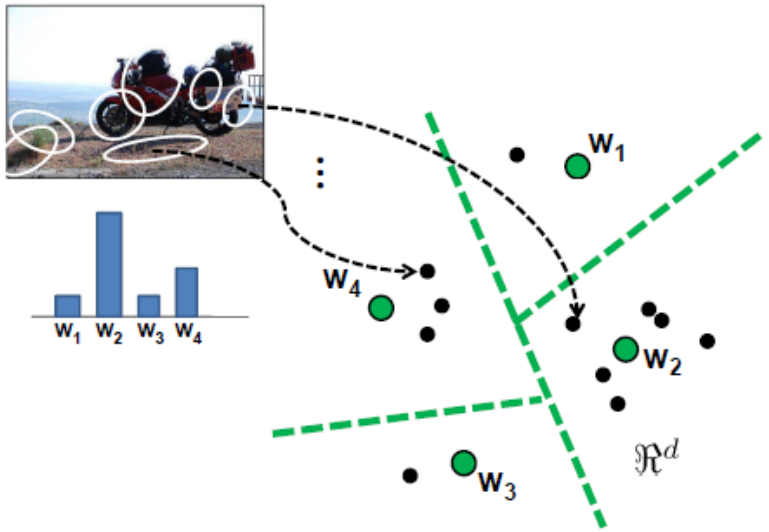
- More advanced quantization / encoding methods that represent the state-of-the-art in image classification and image retrieval.
  - Soft assignment (a.k.a. Kernel Codebook)
  - VLAD
  - Locally Linear Coding

- Classification based on Latent Variables

Motivation

Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**?



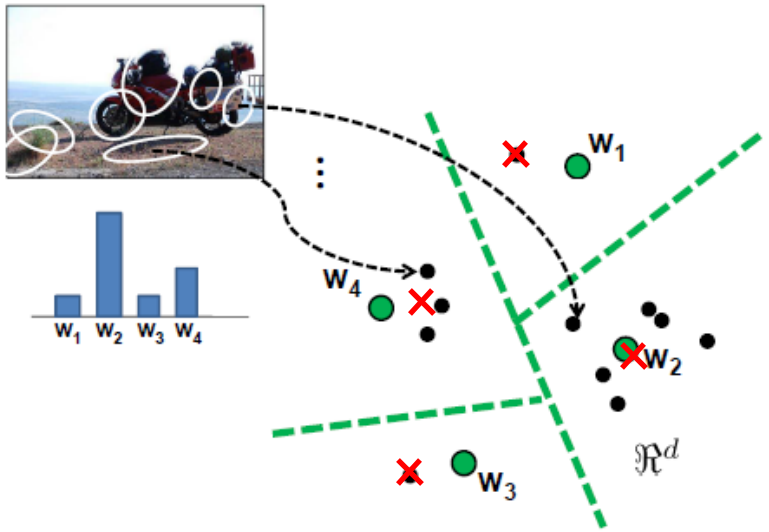
http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Motivation

Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not include **other statistics**? For instance:

- mean of local descriptors ×



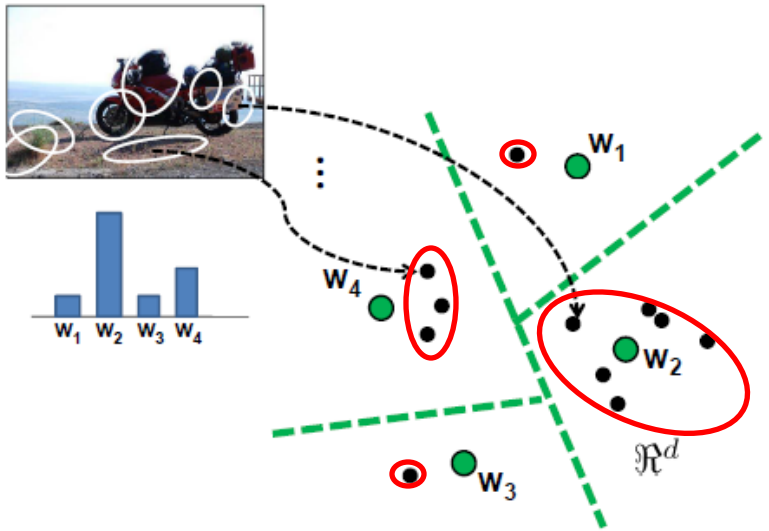
http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Motivation

Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**? For instance:

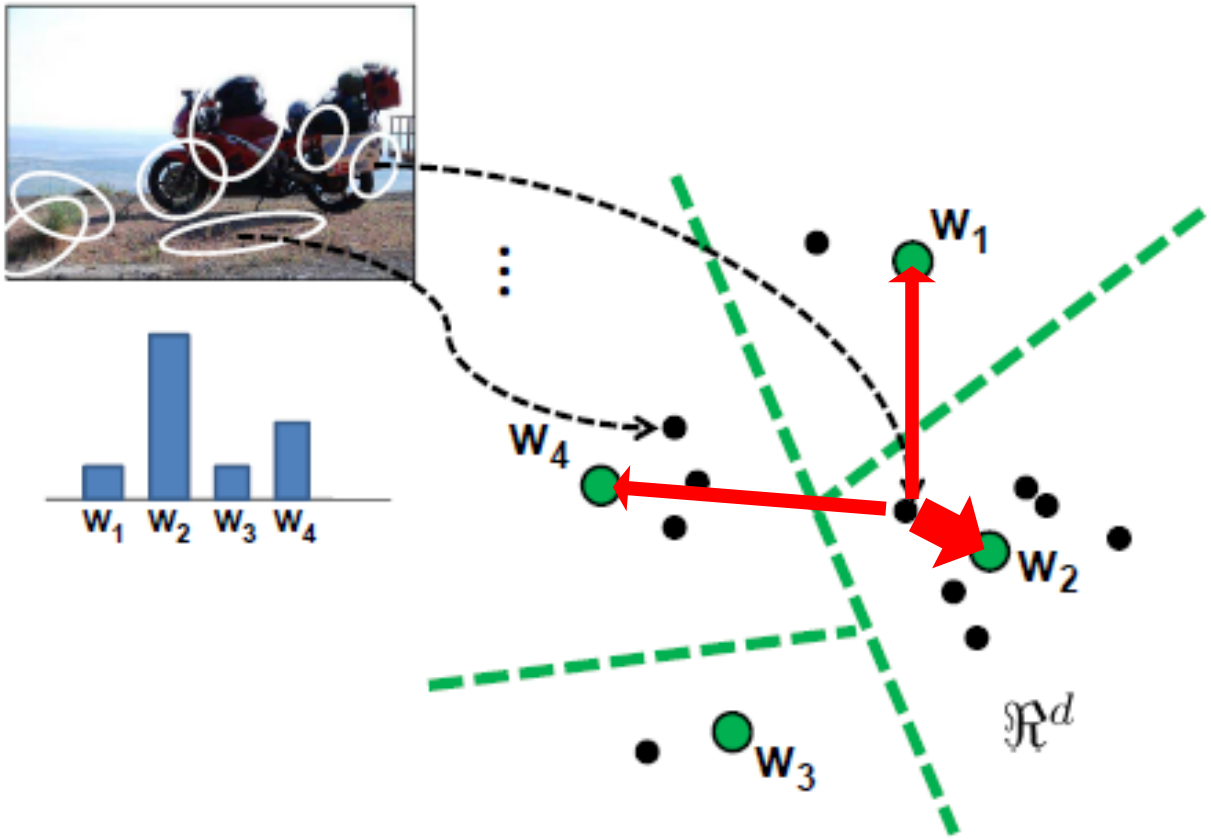
- mean of local descriptors
- (co)variance of local descriptors



http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Simple case: Soft Assignment

- Called “Kernel codebook encoding” by Chatfield et al. 2011. Cast a weighted vote into the most similar clusters.

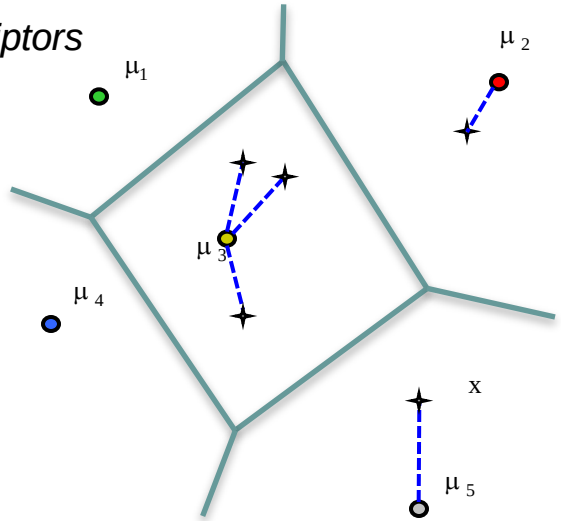


The VLAD representation

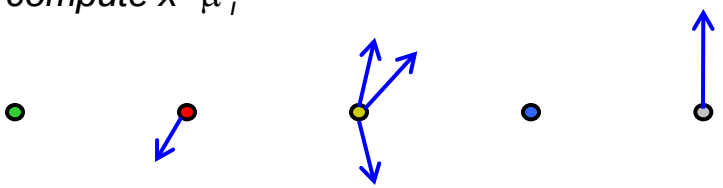
Given a codebook $\{\mu_i, i = 1 \dots N\}$, e.g. learned with K-means, and a set of local descriptors $X = \{x_t, t = 1 \dots T\}$

- ① assign $\mathrm{NN}(x_t) = \arg\min_{\mu_i} ||x_t - \mu_i||$
- ②③ compute: $v_i = \sum_{x_t : \mathrm{NN}(x_t) = \mu_i} x_t - \mu_i$
- concatenate v_i's + ℓ_2 normalize

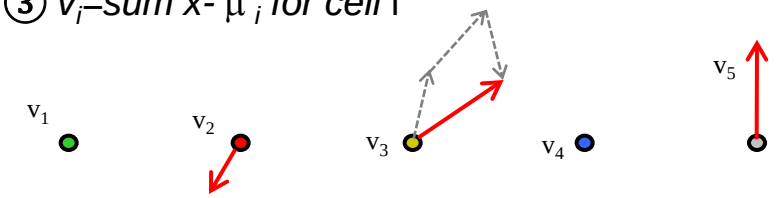
① *assign descriptors* :



② *compute x-* μ_i



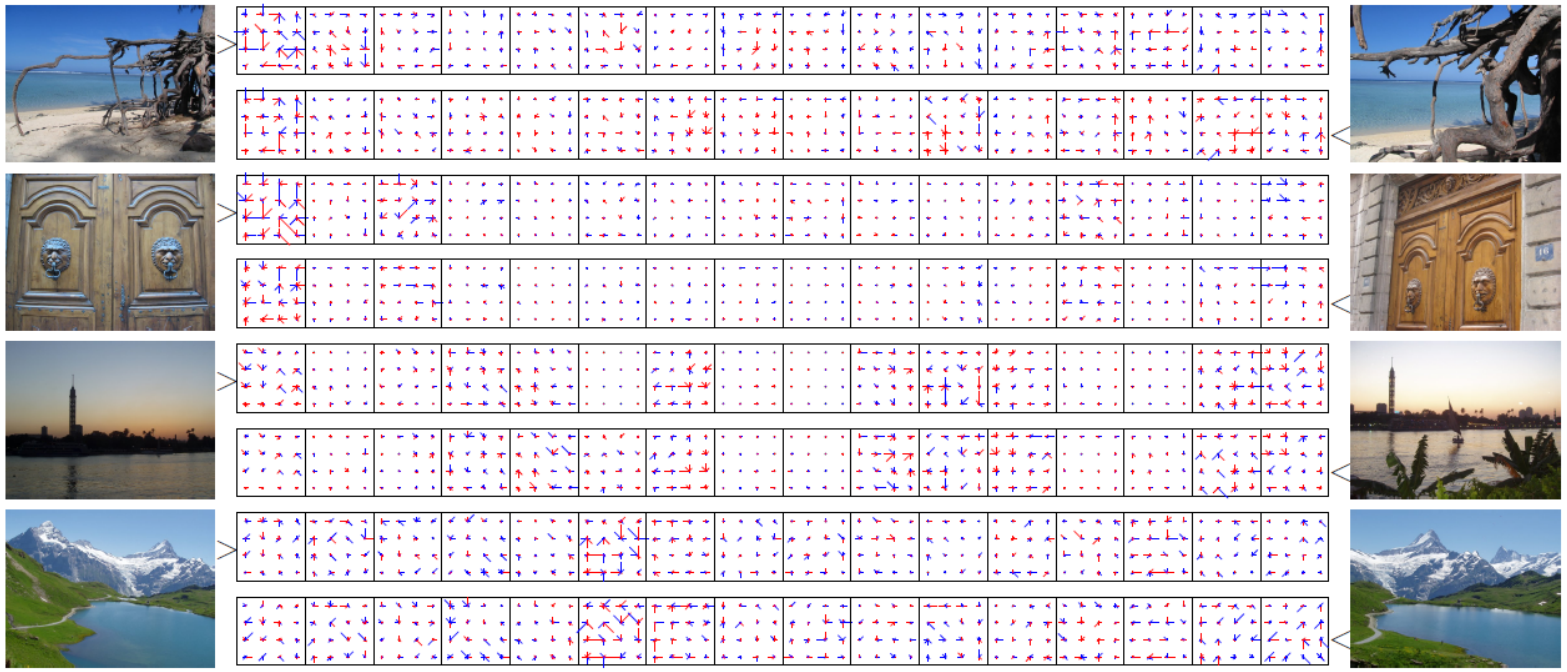
③ *v_i=sum x-* μ_i *for cell* i



Jégou, Douze, Schmid and Pérez, "Aggregating local descriptors into a compact image representation", CVPR'10.

A first example: the VLAD

A graphical representation of $v_i = \sum_{x_t : \mathrm{NN}(x_t) = \mu_i} x_t - \mu_i$



Jégou, Douze, Schmid and Pérez, “Aggregating local descriptors into a compact image representation”, CVPR’10.

Local Coordinate Coding (LCC)

Yu, Zhang & Gong, NIPS 09
Wang, Yang, Yu, Lv, Huang CVPR 10

- Dictionary Learning: k-means (or hierarchical k-means)

- Coding for x, to obtain its sparse representation *a*

  Step 1 – ensure locality: find the K nearest bases

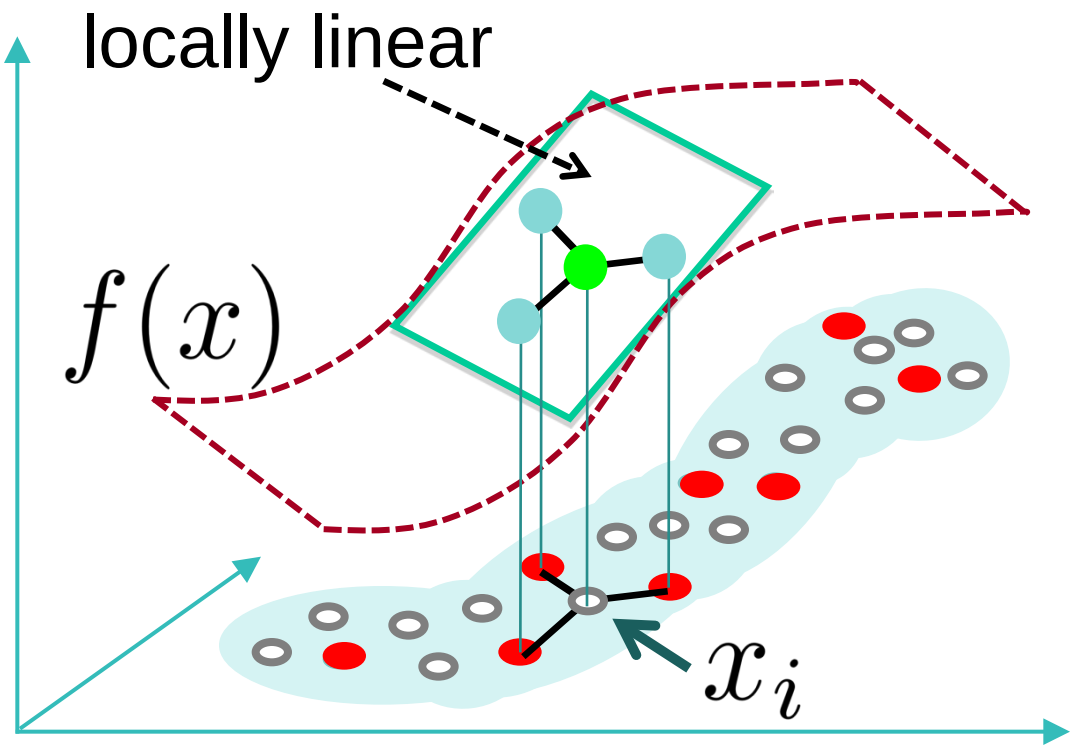
$$[\phi_j]_{j \in J(x)}$$

  Step 2 – ensure low coding error:

$$\min_a \left\| x - \sum_{j \in J(x)} a_{i,j} \phi_j \right\|^2, \quad \text{s.t.} \quad \sum_{j \in J(x)} a_{i,j} = 1$$

Function Approximation based on LCC

Yu, Zhang, Gong, NIPS 10



xerox

KU LEUVEN

PSI VISICS

Object Classification with Latent Variables

IJCV 2013, BMVC 2011

H. Bilen, V.P. Namboodiri and L.V. Gool

VISICS, ESAT, KU Leuven

November 19, 2013

Outline

Visual Object Recognition

?

Visual Object Recognition

?

Visual Object Recognition

observed data

?

Visual Object Recognition

observed data

LABEL

Jean Claude Van Damme

Visual Object Recognition

observed data

LABEL

Jean Claude Van Damme

- instance specific recognition

Visual Object Recognition

observed data

Jean Claude Van Damme

LABEL

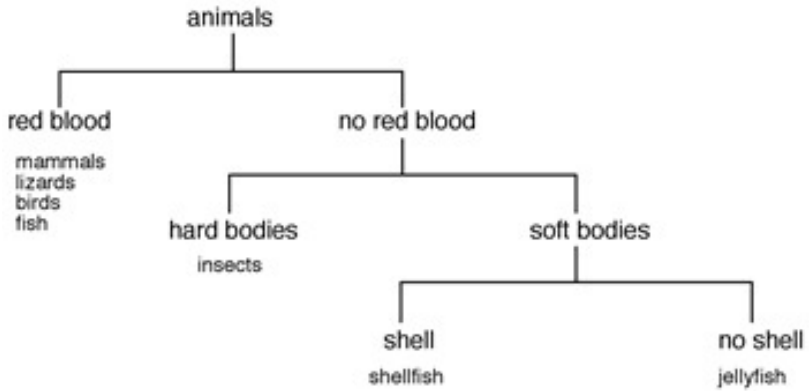
- instance specific recognition
- generic class recognition *e.g.* bicycle, person

What is an object class?

What is an object class?

Classical view

- members share a set of properties



What is an object class?

Classical view

- members share a set of properties
- finding the common
- functional or visual?

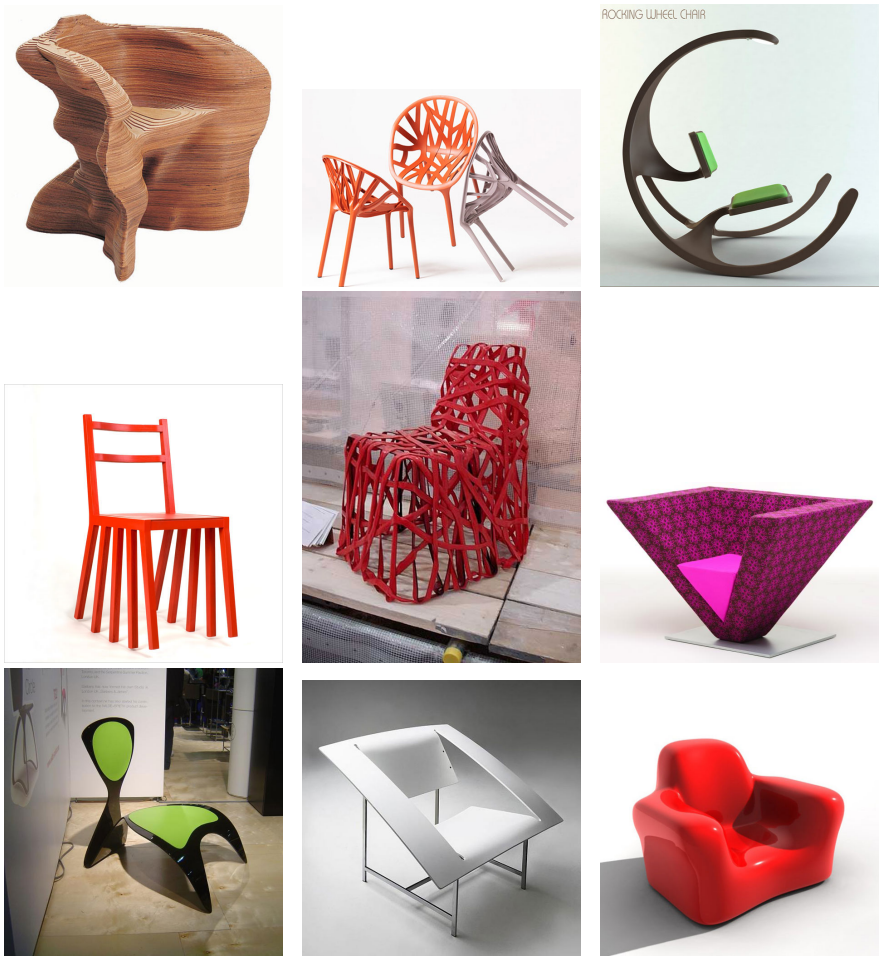


image credits: webdesignerdepot.com

What is an object class?

Classical view

- members share a set of properties
- finding the common
- functional or visual?

Exemplar view

- shared properties $\Rightarrow$ *similarity*

image credits: PASCAL VOC 2007 dataset

What is an object class?

Classical view

- members share a set of properties
- finding the common
- functional or visual?

Exemplar view

- shared properties $\Rightarrow$ *similarity*

image credits: PASCAL VOC 2007 dataset

What is an object class?

Classical view

- members share a set of properties
- finding the common
- functional or visual?

Exemplar view

- shared properties $\Rightarrow$ *similarity*
- difficult to evaluate
- requires a huge collection

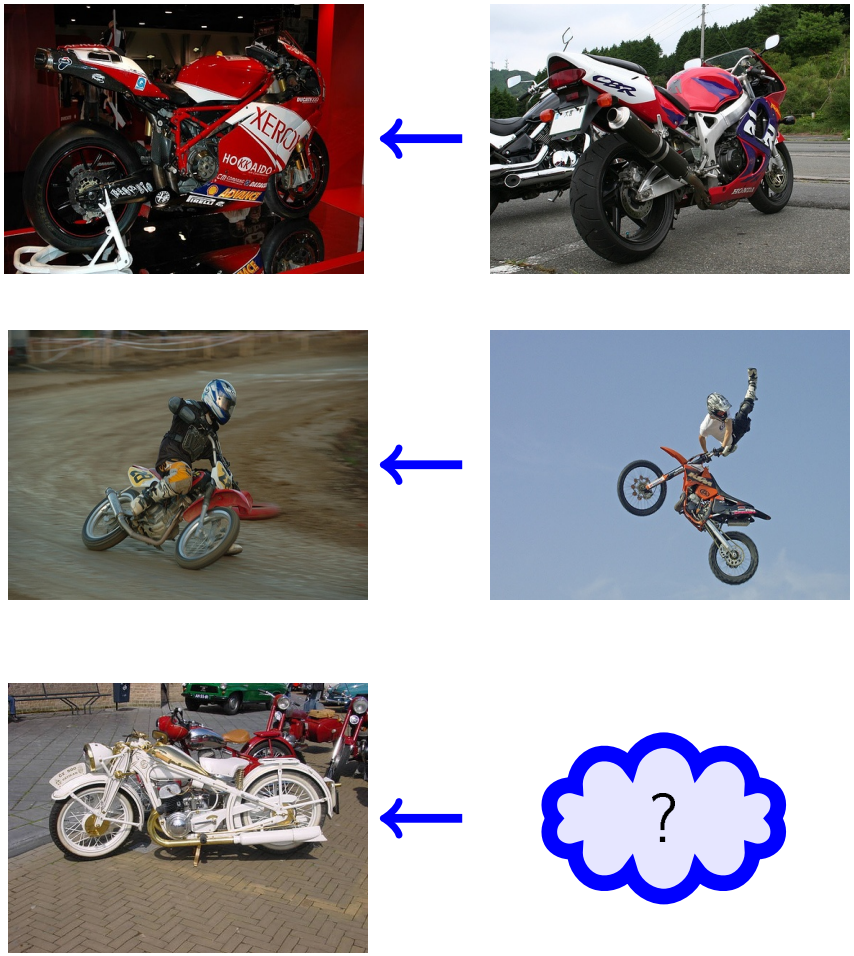


image credits: PASCAL VOC 2007 dataset

What is an object class?

Classical view

- members share a set of properties
- finding the common
- functional or visual?

Exemplar view

- shared properties $\Rightarrow$ *similarity*
- difficult to evaluate
- requires a huge collection

In practice

- closed, well defined sets
- train and test sets

motorbike

not motorbike

What is an object class?

Classical view

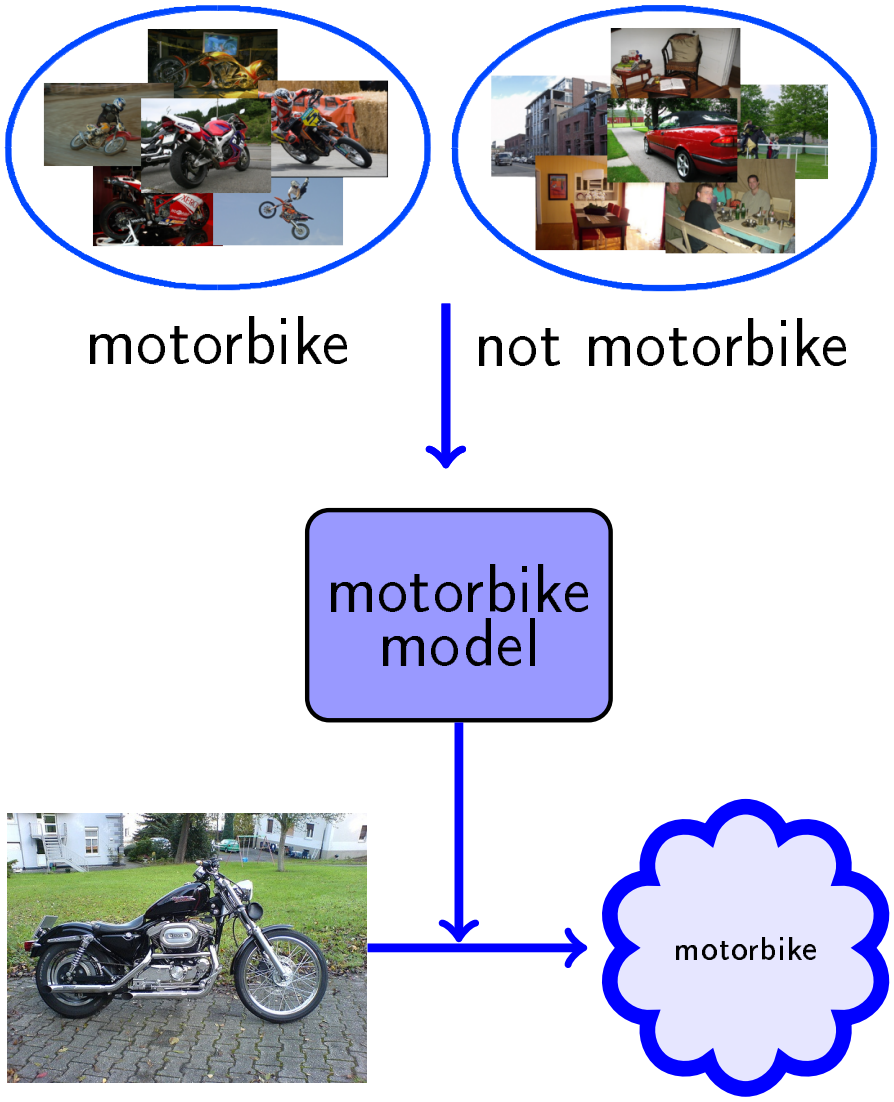
- members share a set of properties
- finding the common
- functional or visual?

Exemplar view

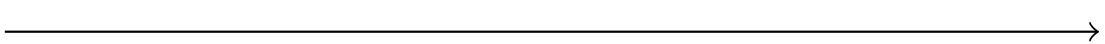
- shared properties $\Rightarrow$ *similarity*
- difficult to evaluate
- requires a huge collection

In practice

- closed, well defined sets
- train and test sets

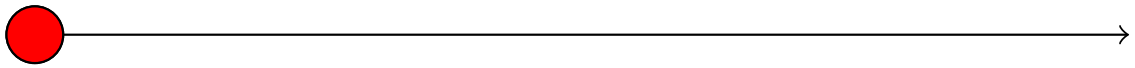


Recognition and Localization



localization

Recognition and Localization

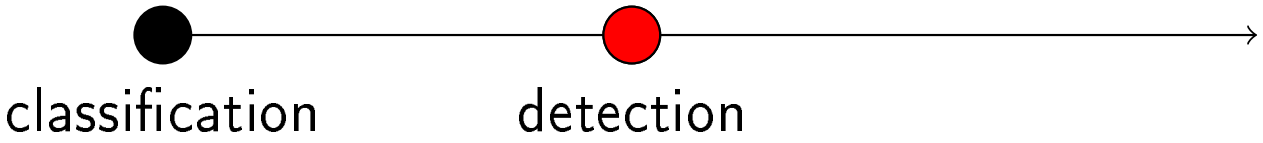


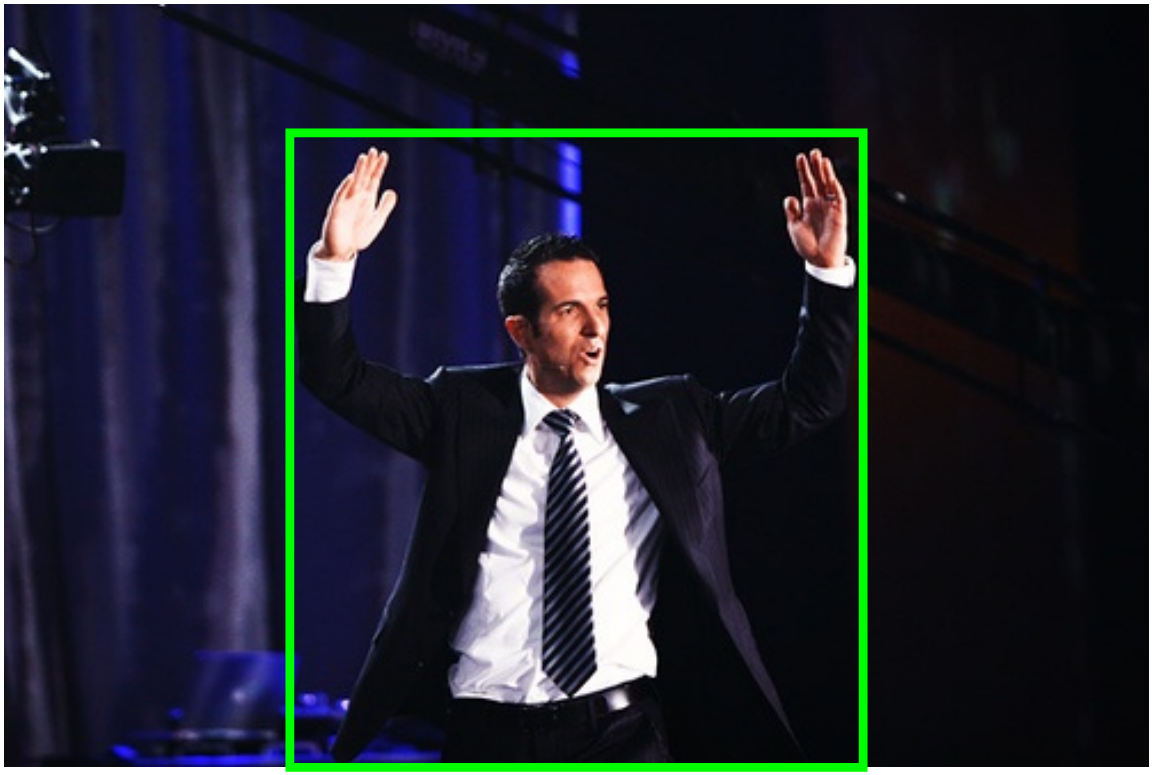
classification

localization

label

Recognition and Localization





bounding box and *label*

Recognition and Localization

label at each pixel

Recognition and Localization

label

Why does it matter?

Why does it matter?

Intellectual curiosity

- Can we extract semantic information from a signal?
- Can we make machines see?

Why does it matter?

Intellectual curiosity

- Can we extract semantic information from a signal?
- Can we make machines see?

Applications

- Assistive technologies
- Semantic image and video search
- Surveillance
- Medical imaging
- Human-computer interaction
- ...

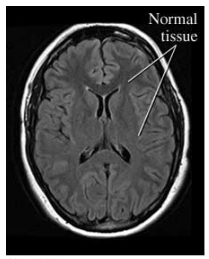


Figure 1

Figure 2

image credits: blog.timesunion.com, renttycoons.com, Intermountain Medical Imaging, digitaltrends.com, iwatchsystems.com

Challenges in Classification

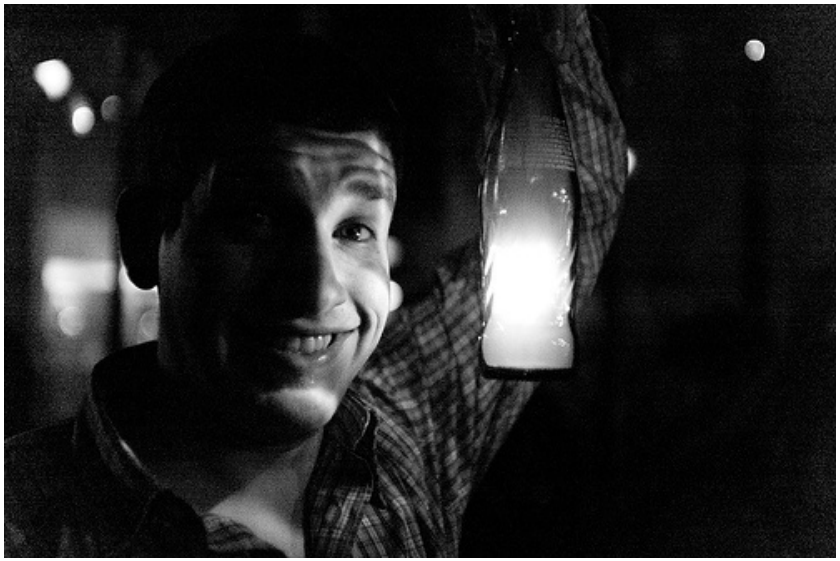
Sources of variation

- illumination
- viewpoint
- background clutter
- occlusion
- intra-class variability
- intra-class vs. inter-class variability

Challenges in Classification

Sources of variation

- illumination
- viewpoint
- background clutter
- occlusion
- intra-class variability
- intra-class vs. inter-class variability



Challenges in Classification

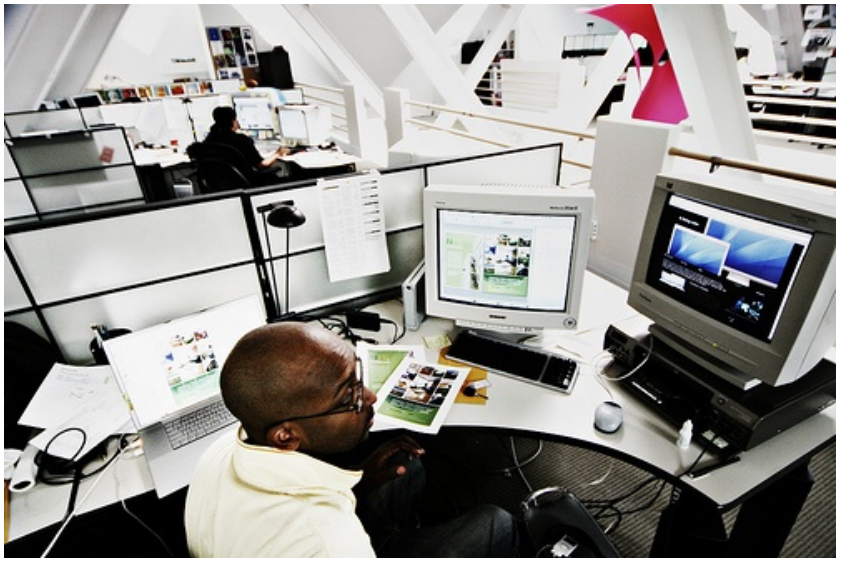
Sources of variation

- illumination
- viewpoint
- background clutter
- occlusion
- intra-class variability
- intra-class vs. inter-class variability

Challenges in Classification

Sources of variation

- illumination
- viewpoint
- background clutter
- occlusion
- intra-class variability
- intra-class vs. inter-class variability



Challenges in Classification

Sources of variation

- illumination
- viewpoint
- background clutter
- occlusion
- intra-class variability
- intra-class vs. inter-class variability

Challenges in Classification

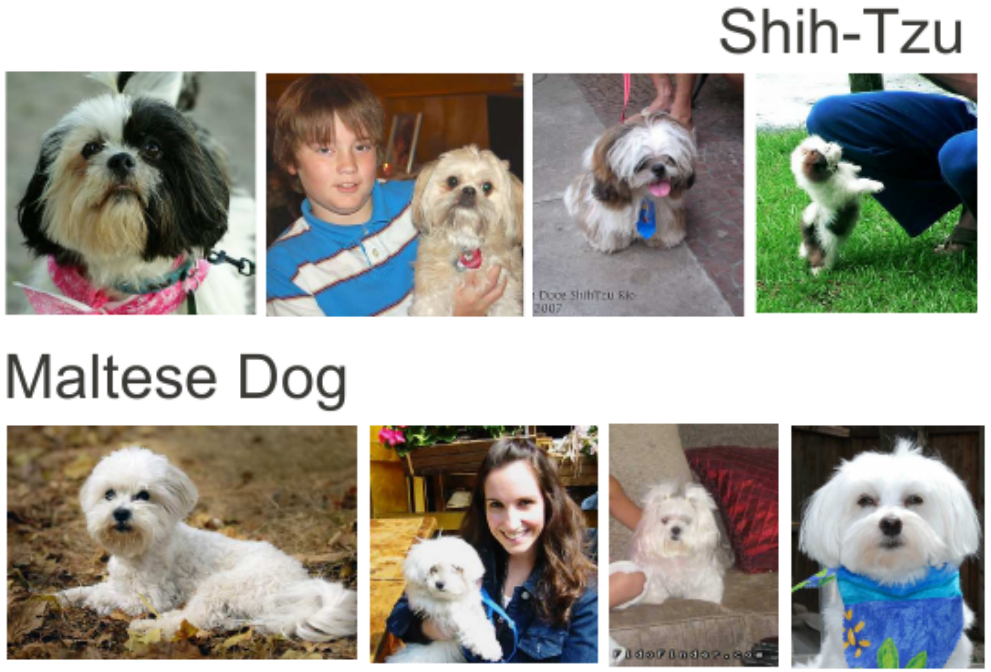
Sources of variation

- illumination
- viewpoint
- background clutter
- occlusion
- intra-class variability
- intra-class vs. inter-class variability

Challenges in Classification

Sources of variation

- illumination
- viewpoint
- background clutter
- occlusion
- intra-class variability
- intra-class vs. inter-class variability



Coping with the challenges

develop invariant representations

Coping with the challenges

develop invariant representations

- pros: no extra parameters
- cons: less discriminative

Coping with the challenges

develop invariant representations

- pros: no extra parameters
- cons: less discriminative

model variances explicitly

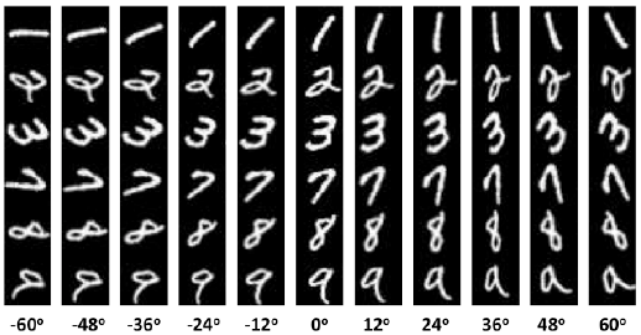
Coping with the challenges

develop invariant representations

- pros: no extra parameters
- cons: less discriminative

model variances explicitly

- pros: more discriminative
- cons:
  - extra parameters



Coping with the challenges

develop invariant representations

- pros: no extra parameters
- cons: less discriminative

model variances explicitly

- pros: more discriminative
- cons:
  - extra parameters
  - annotation

Annotating Images

- time-consuming, expensive, not much fun!

image credits: Barriosa and Torralba

Annotating Images

- time-consuming, expensive, not much fun!

image credits: Barriosa and Torralba

Goal

improve classification task by

Goal

improve classification task by

- modeling various sources of variation,

Goal

improve classification task by

- modeling various sources of variation,
- without requiring any annotation but only class labels,

Goal

improve classification task by

- modeling various sources of variation,
- without requiring any annotation but only class labels,
- incorporating them into image representation as latent variables,

Goal

improve classification task by

- modeling various sources of variation,
- without requiring any annotation but only class labels,
- incorporating them into image representation as latent variables,
- jointly learning both classifier parameters and latent parameters in a discriminative setting

Previous Work

Object Classification

- state of the art classification methods

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]
  - multiple features [Gehler@ICCV09], classifiers [Torresani@ECCV10]

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]
  - multiple features [Gehler@ICCV09], classifiers [Torresani@ECCV10]
- static representations

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]
  - multiple features [Gehler@ICCV09], classifiers [Torresani@ECCV10]
- static representations
- various quantization and encoding methods are compatible with our work

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]
  - multiple features [Gehler@ICCV09], classifiers [Torresani@ECCV10]
- static representations
- various quantization and encoding methods are compatible with our work

Latent Variable Models

- generative methods (GMMs, HMMs)

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]
  - multiple features [Gehler@ICCV09], classifiers [Torresani@ECCV10]
- static representations
- various quantization and encoding methods are compatible with our work

Latent Variable Models

- generative methods (GMMs, HMMs)
- discriminative methods (Hidden CRFs [Wang@CVPR06], Latent SVM [Felzenswalb@PAMI10, Yu@ICML09])

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]
  - multiple features [Gehler@ICCV09], classifiers [Torresani@ECCV10]
- static representations
- various quantization and encoding methods are compatible with our work

Latent Variable Models

- generative methods (GMMs, HMMs)
- discriminative methods (Hidden CRFs [Wang@CVPR06], Latent SVM [Felzenswalb@PAMI10, Yu@ICML09])
- in both cases parameter learning is non-convex

Previous Work

Object Classification

- state of the art classification methods
  - better feature quantization and encoding, LLC [Wang@CVPR09]
  - more powerful statistics, I. Fisher Encoding [Perronnin@ECCV10]
  - multiple features [Gehler@ICCV09], classifiers [Torresani@ECCV10]
- static representations
- various quantization and encoding methods are compatible with our work

Latent Variable Models

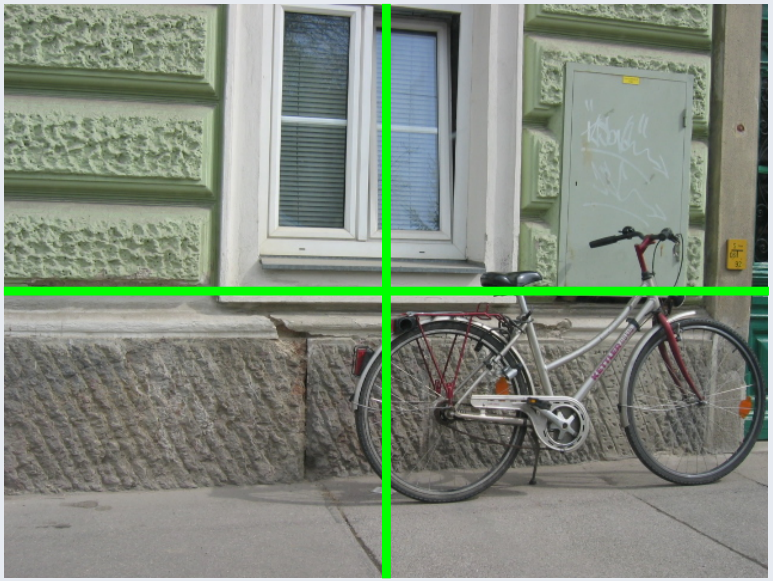
- generative methods (GMMs, HMMs)
- discriminative methods (Hidden CRFs [Wang@CVPR06], Latent SVM [Felzenswalb@PAMI10, Yu@ICML09])
- in both cases parameter learning is non-convex
  - EM (generative)
  - CCCP (discriminative)

Learning Spatial Pyramids

Spatial Pyramids [Lazebnik@CVPR06]

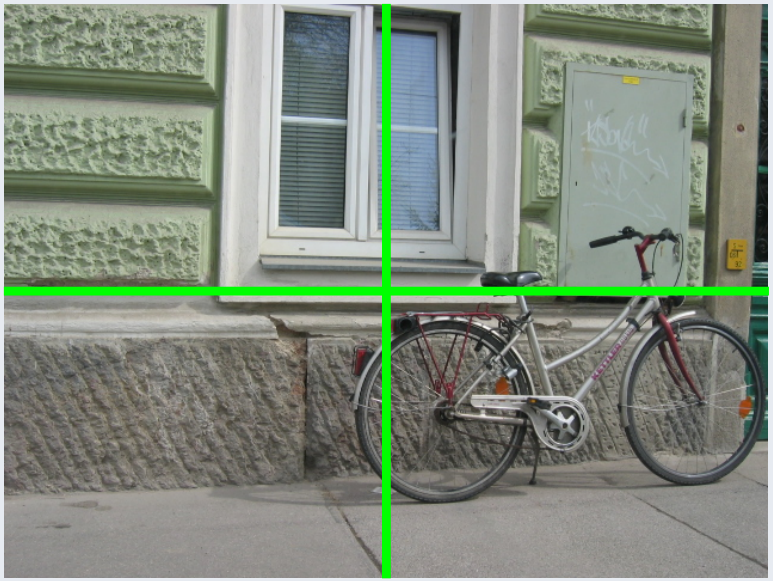
Learning Spatial Pyramids

Spatial Pyramids [Lazebnik@CVPR06]



Learning Spatial Pyramids

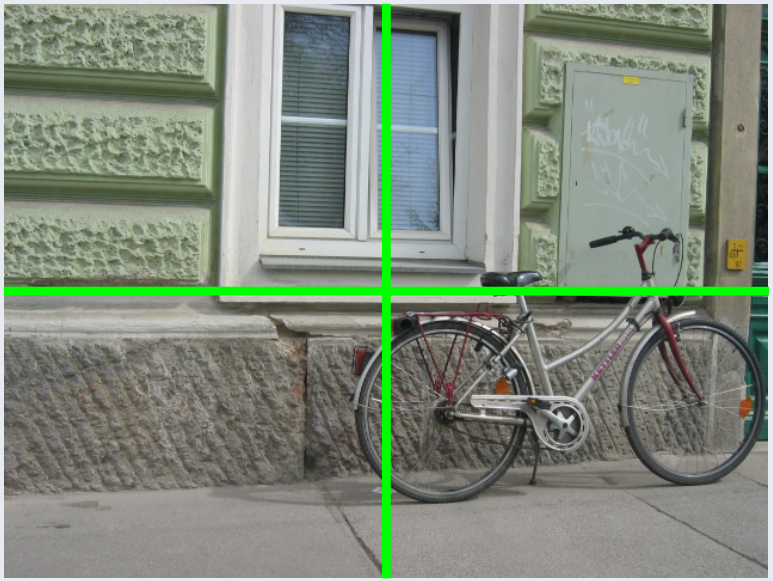
Spatial Pyramids [Lazebnik@CVPR06]



- pros: adds spatial information to the orderless BoW

Learning Spatial Pyramids

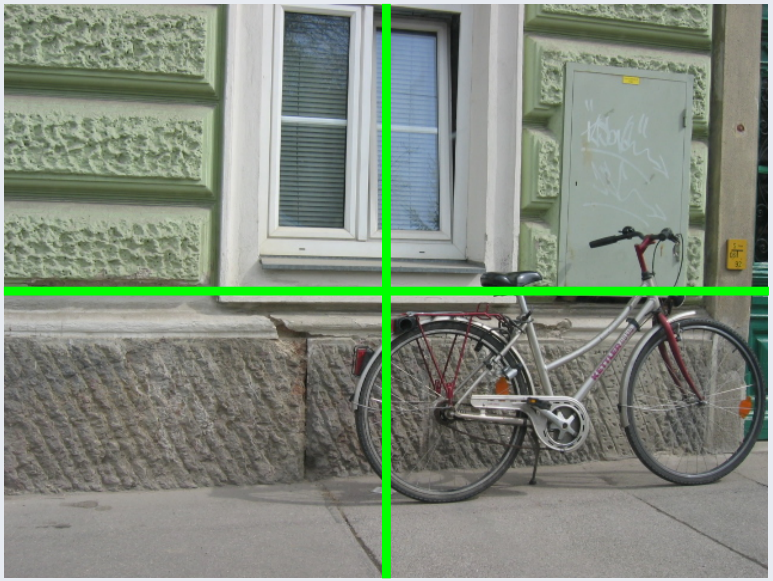
Spatial Pyramids [Lazebnik@CVPR06]



- pros: adds spatial information to the orderless BoW
- cons:

Learning Spatial Pyramids

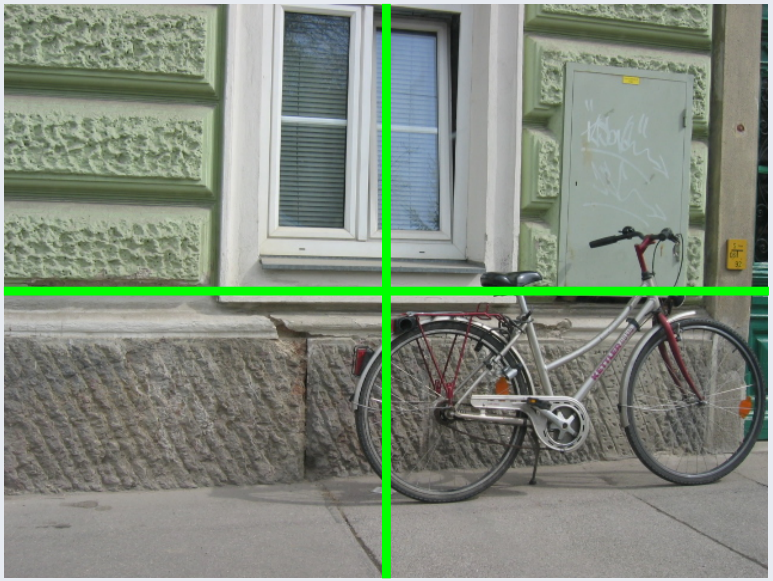
Spatial Pyramids [Lazebnik@CVPR06]



- pros: adds spatial information to the orderless BoW
- cons:
  - static, not optimal

Learning Spatial Pyramids

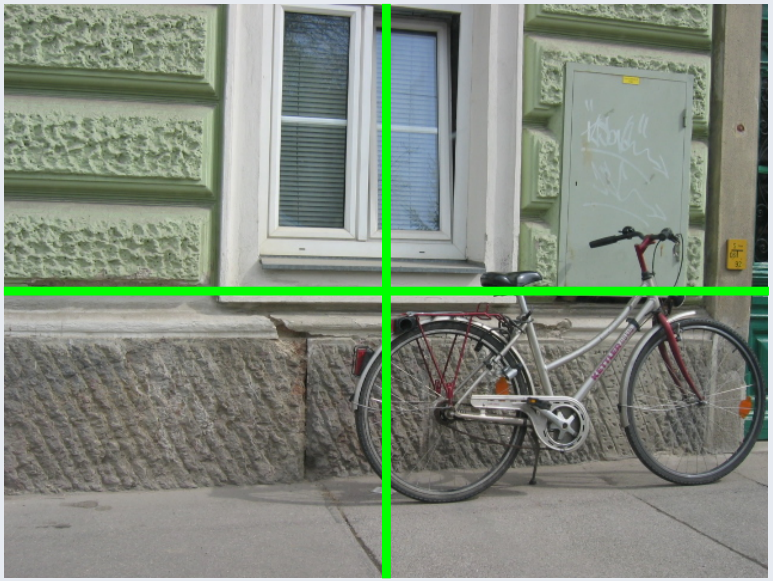
Spatial Pyramids [Lazebnik@CVPR06]



- pros: adds spatial information to the orderless BoW
- cons:
  - static, not optimal
  - features in cells are not aligned between images

Learning Spatial Pyramids

Spatial Pyramids [Lazebnik@CVPR06]



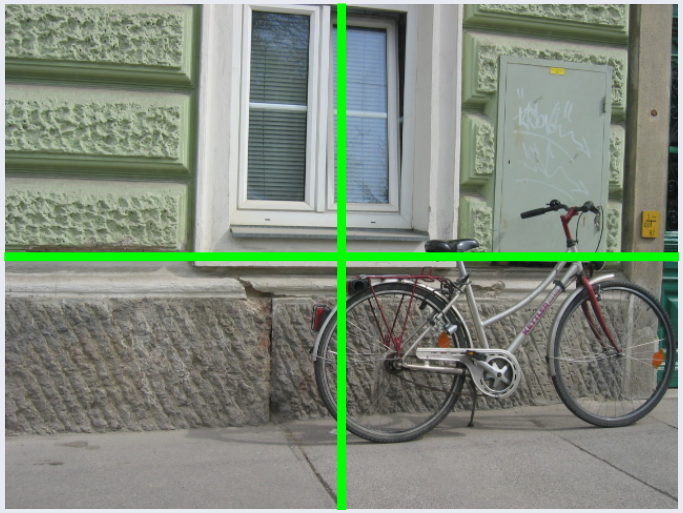
- pros: adds spatial information to the orderless BoW
- cons:
  - static, not optimal
  - features in cells are not aligned between images
  - includes discriminative and non-representative features

More Flexible Image Representation

Split

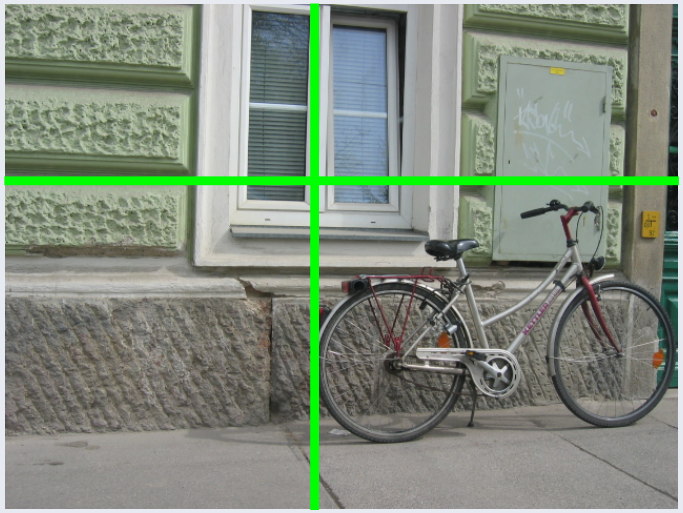
More Flexible Image Representation

Split



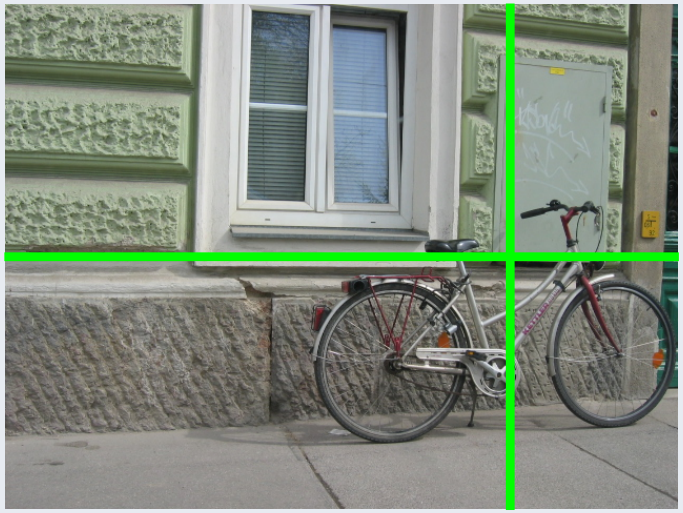
More Flexible Image Representation

Split



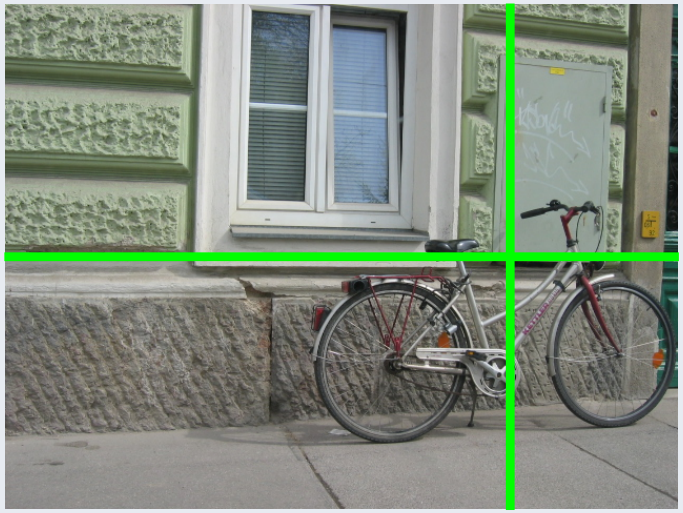
More Flexible Image Representation

Split

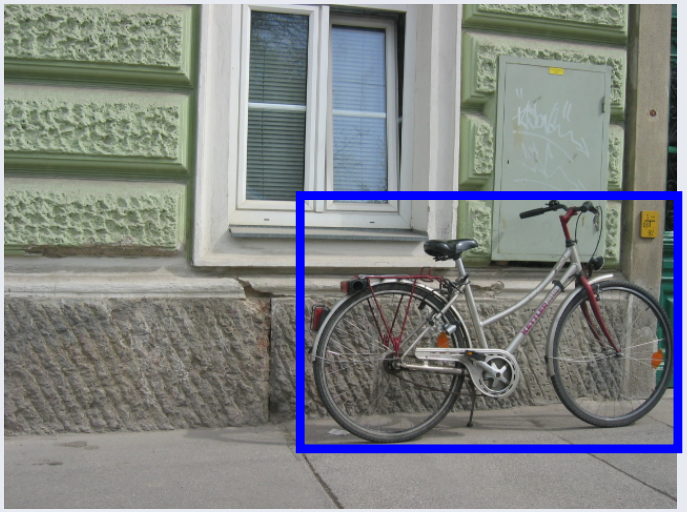


More Flexible Image Representation

Split

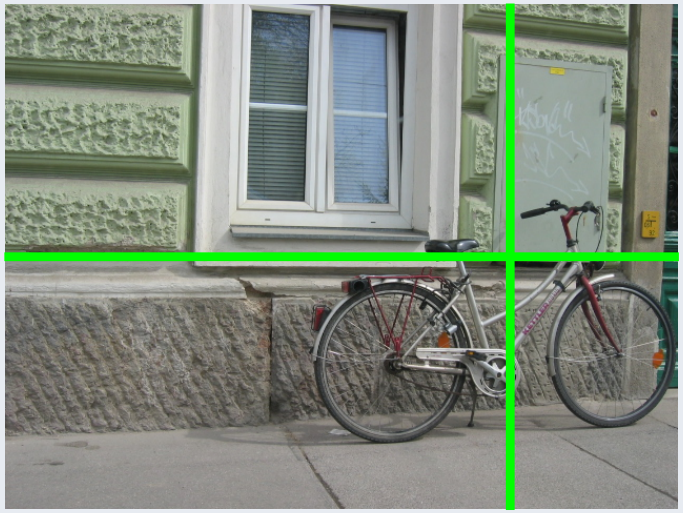


Crop

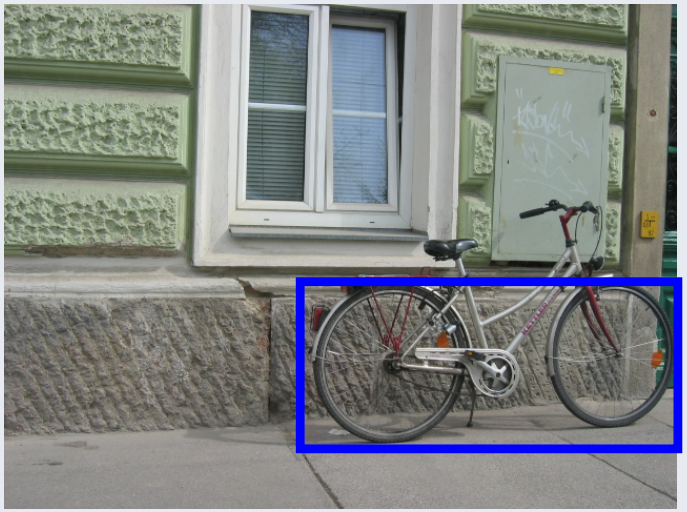


More Flexible Image Representation

Split

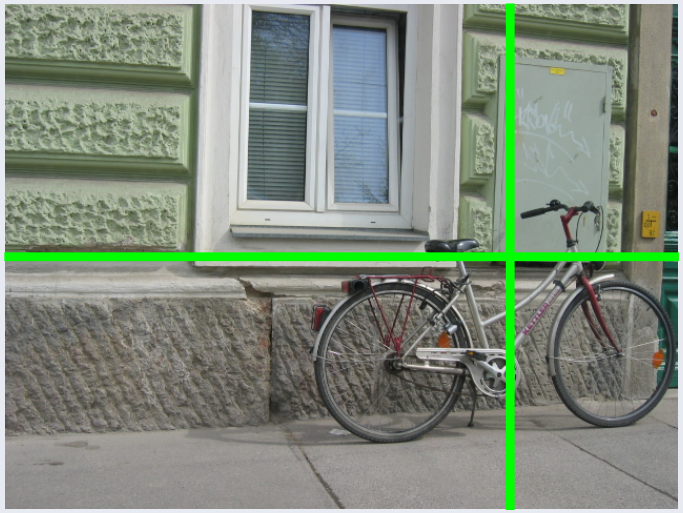


Crop

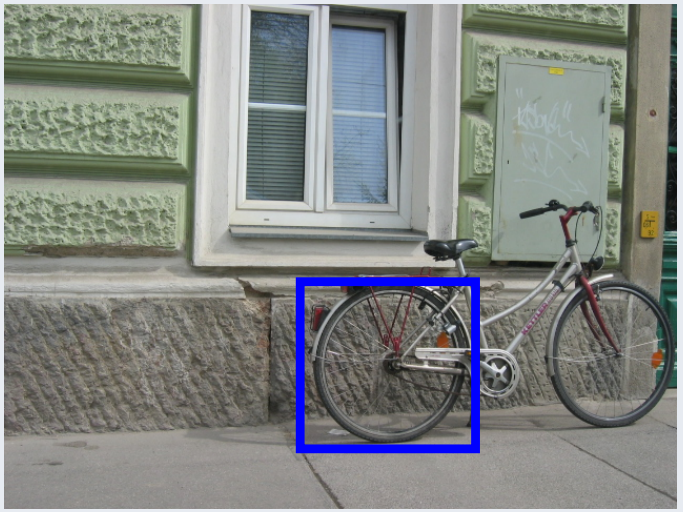


More Flexible Image Representation

Split

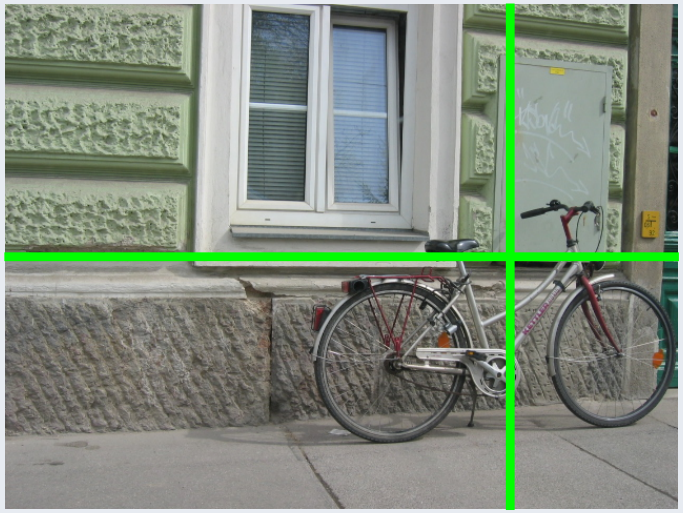


Crop

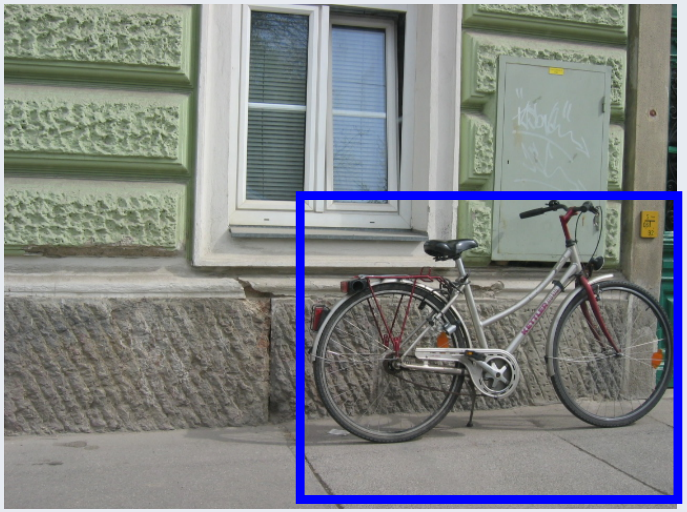


More Flexible Image Representation

Split

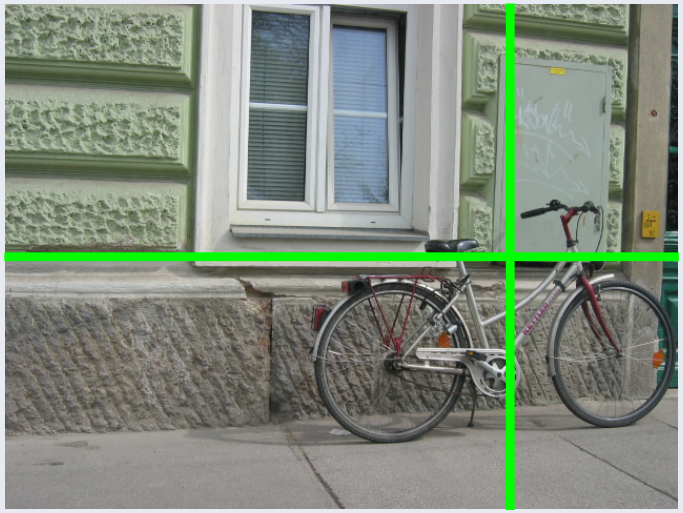


Crop

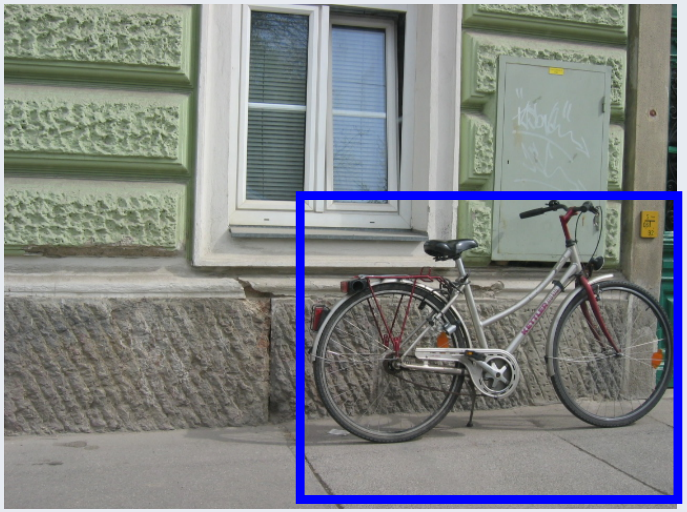


More Flexible Image Representation

Split



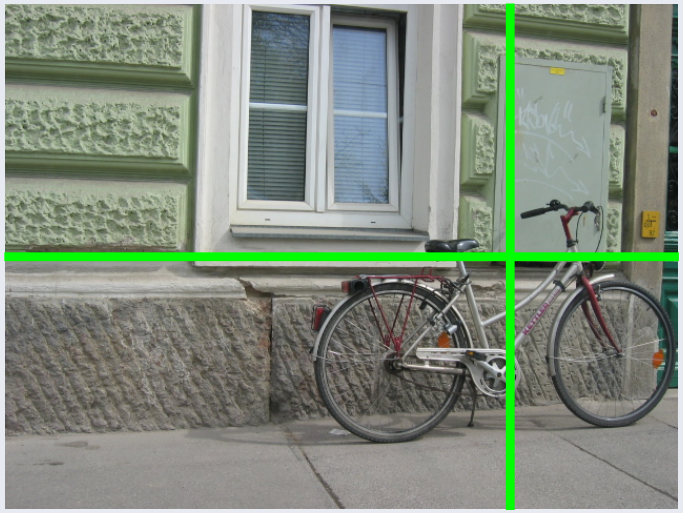
Crop



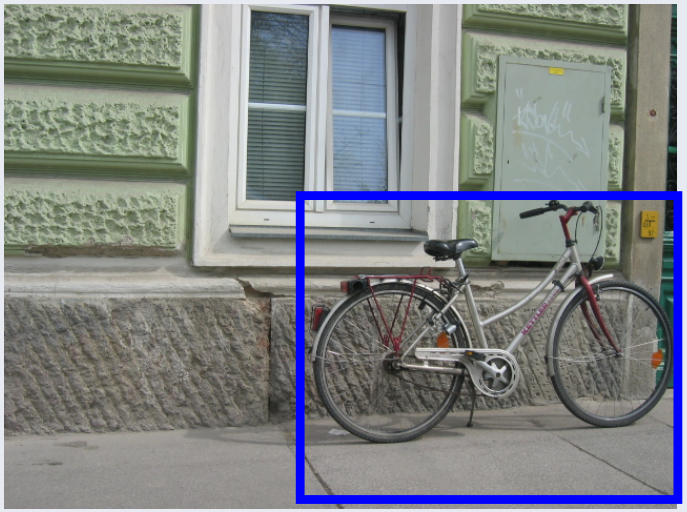
Crop - Uniform Split

More Flexible Image Representation

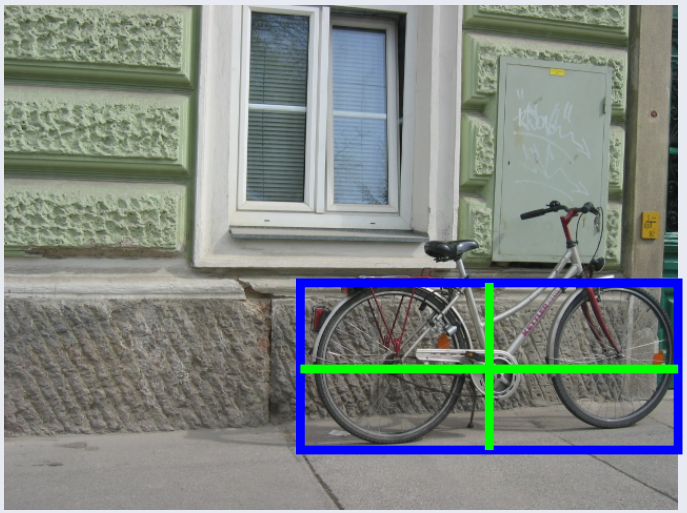
Split



Crop

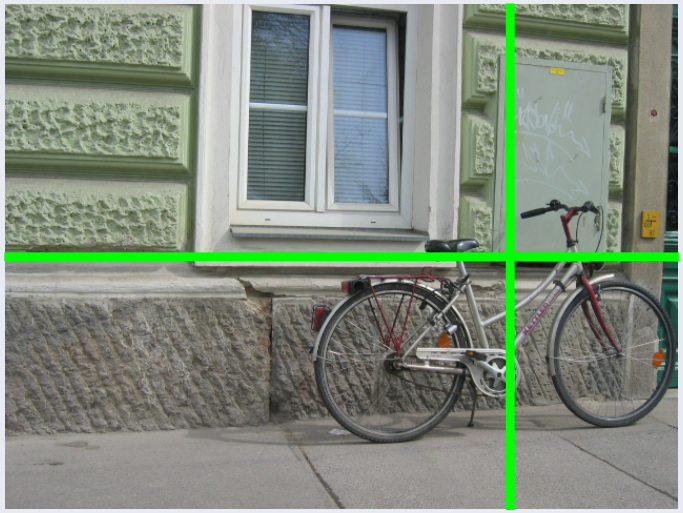


Crop - Uniform Split

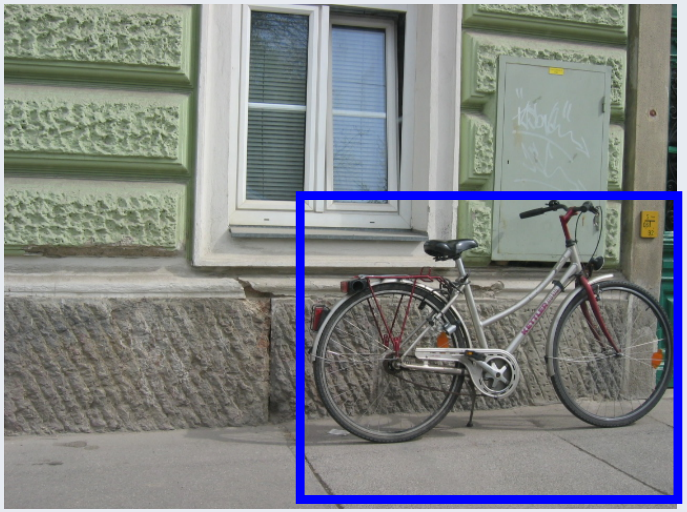


More Flexible Image Representation

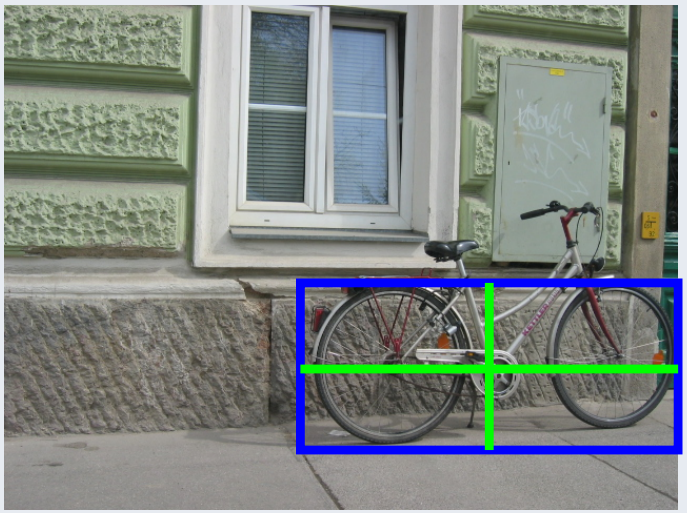
Split



Crop



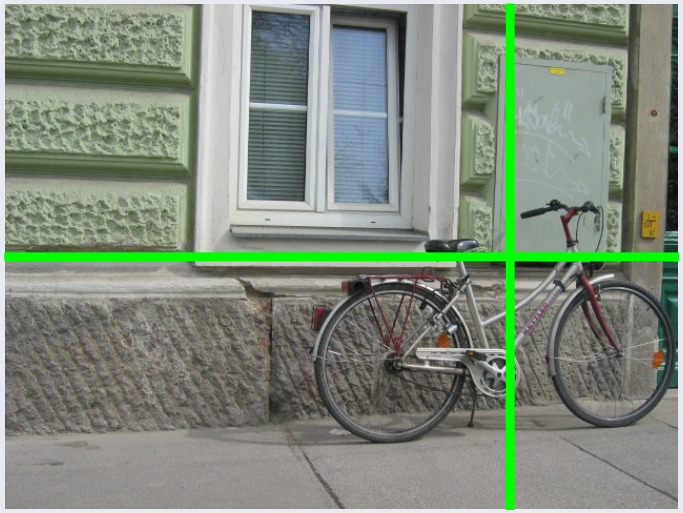
Crop - Uniform Split



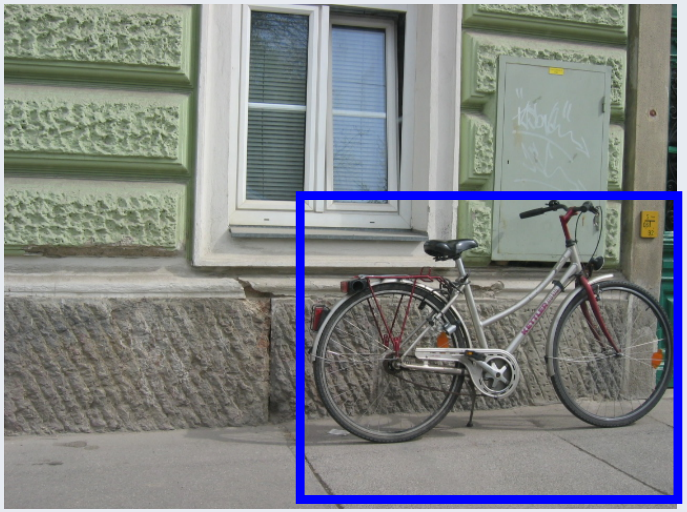
Crop - Split

More Flexible Image Representation

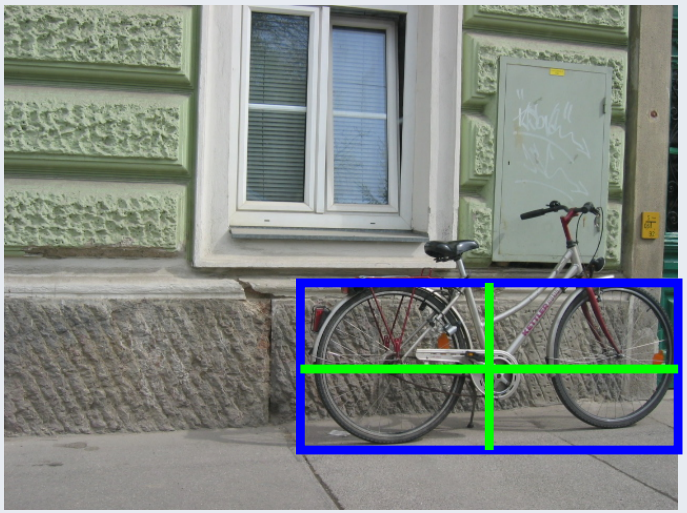
Split



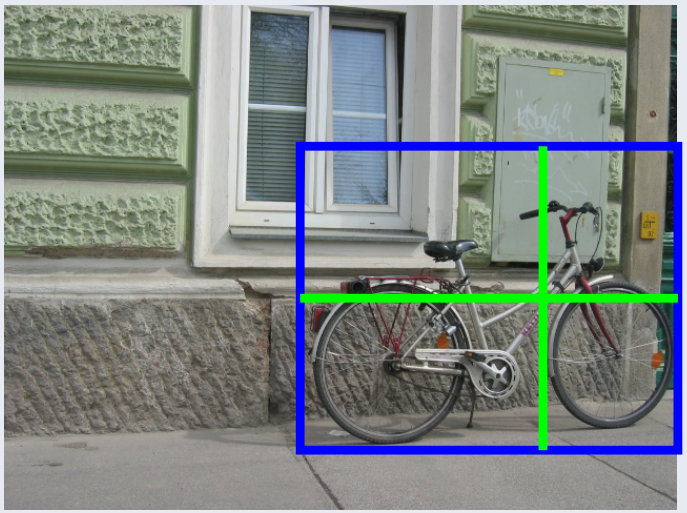
Crop



Crop - Uniform Split



Crop - Split



Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks

Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks
- disjoint learning is not optimal in terms of classification performance

Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks
- disjoint learning is not optimal in terms of classification performance

Latent (Structural) SVM [Yu@ICML09]

Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks
- disjoint learning is not optimal in terms of classification performance

Latent (Structural) SVM [Yu@ICML09]

- image windows/non-uniform grid (crop/split) as latent parameters

Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks
- disjoint learning is not optimal in terms of classification performance

Latent (Structural) SVM [Yu@ICML09]

- image windows/non-uniform grid (crop/split) as latent parameters
- alternating optimization between classification parameters and latent parameters

Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks
- disjoint learning is not optimal in terms of classification performance

Latent (Structural) SVM [Yu@ICML09]

- image windows/non-uniform grid (crop/split) as latent parameters
- alternating optimization between classification parameters and latent parameters
- is non-convex and requires initialization of classification or latent parameters

Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks
- disjoint learning is not optimal in terms of classification performance

Latent (Structural) SVM [Yu@ICML09]

- image windows/non-uniform grid (crop/split) as latent parameters
- alternating optimization between classification parameters and latent parameters
- is non-convex and requires initialization of classification or latent parameters
- is guaranteed to decrease the objective function and to converge to a local minimum

Learning to Localize and to Classify

- learning classifier parameters and image windows/non-uniform grid are dependent tasks
- disjoint learning is not optimal in terms of classification performance

Latent (Structural) SVM [Yu@ICML09]

- image windows/non-uniform grid (crop/split) as latent parameters
- alternating optimization between classification parameters and latent parameters
- is non-convex and requires initialization of classification or latent parameters
- is guaranteed to decrease the objective function and to converge to a local minimum
- We use Latent SVM for learning

Latent Structural SVM Formulation

Training pairs $\{(\mathbf{x_i}, y_i, \mathbf{h_i})\} \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$

- $\mathbf{x_i}$ are images/videos, y_i are class labels,
- $\mathbf{h_i}$ are coordinates of image windows/non-uniform grids.
- only $\mathbf{x_i}$ and y_i are given, $\mathbf{h_i}$ are not given

Latent Structural SVM Formulation

Training pairs $\{(\mathbf{x_i}, \mathbf{y_i}, \mathbf{h_i})\} \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$

- $\mathbf{x_i}$ are images/videos, $\mathbf{y_i}$ are class labels,
- $\mathbf{h_i}$ are coordinates of image windows/non-uniform grids.
- only $\mathbf{x_i}$ and $\mathbf{y_i}$ are given, $\mathbf{h_i}$ are not given

Discriminant function that measures the matching quality of the triplet:

$$\mathbf{f} : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \rightarrow \mathcal{R}$$

$$\mathbf{f_w}(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\psi}(\mathbf{x}, \mathbf{y}, \mathbf{h})$$

Latent Structural SVM Formulation

Training pairs $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i)\} \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$

- $\mathbf{x}_i$ are images/videos, $\mathbf{y}_i$ are class labels,
- $\mathbf{h}_i$ are coordinates of image windows/non-uniform grids.
- only $\mathbf{x}_i$ and $\mathbf{y}_i$ are given, $\mathbf{h}_i$ are not given

Discriminant function that measures the matching quality of the triplet:

$$\mathbf{f} : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \rightarrow \mathcal{R}$$

$$\mathbf{f}_{\boldsymbol{w}}(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\psi}(\mathbf{x}, \mathbf{y}, \mathbf{h})$$

- $\boldsymbol{\psi}(\mathbf{x}, \mathbf{y}, \mathbf{h})$ is a histogram of quantized SIFT features which is obtained through the image window or non-uniform grid $\mathbf{h}$.

Latent Structural SVM Formulation

Training pairs $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i)\} \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$

- $\mathbf{x}_i$ are images/videos, $\mathbf{y}_i$ are class labels,
- $\mathbf{h}_i$ are coordinates of image windows/non-uniform grids.
- only $\mathbf{x}_i$ and $\mathbf{y}_i$ are given, $\mathbf{h}_i$ are not given

Discriminant function that measures the matching quality of the triplet:

$$\mathbf{f} : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \rightarrow \mathcal{R}$$

$$\mathbf{f}_{\boldsymbol{w}}(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\psi}(\mathbf{x}, \mathbf{y}, \mathbf{h})$$

- $\boldsymbol{\psi}(\mathbf{x}, \mathbf{y}, \mathbf{h})$ is a histogram of quantized SIFT features which is obtained through the image window or non-uniform grid $\mathbf{h}$.
- $\boldsymbol{w}$ is the parameter vector of the latent SVM.

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x, \hat{y}, \hat{h})}_{\text{prediction}}$$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x,\hat{y},\hat{h})}_{\text{prediction}}$$

$$\min_w \left\{ \underbrace{\Omega(w)}_{\text{regularization}} + C\sum_{i=1}^{n} \left[\underbrace{\max_{\hat{y}_i,\hat{h}_i} \left[f_w(x_i,\hat{y}_i,\hat{h}_i) + \Delta(y_i,\hat{y}_i) \right]}_{\text{most violated constraint}} - \underbrace{\max_{h_i^*} \left[f_w(x_i,y_i,h_i^*) \right]}_{\text{best latent variable}} \right] \right\}$$

- Loss function $\Delta(y_i,\hat{y}_i)$

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x,\hat{y},\hat{h})}_{\text{prediction}}$$

$$\min_w \left\{ \underbrace{\Omega(w)}_{\text{regularization}} + C\sum_{i=1}^{n} \left[\underbrace{\max_{\hat{y}_i,\hat{h}_i} \left[f_w(x_i,\hat{y}_i,\hat{h}_i) + \Delta(y_i,\hat{y}_i) \right]}_{\text{most violated constraint}} - \underbrace{\max_{h_i^*} \left[f_w(x_i,y_i,h_i^*) \right]}_{\text{best latent variable}} \right] \right\}$$

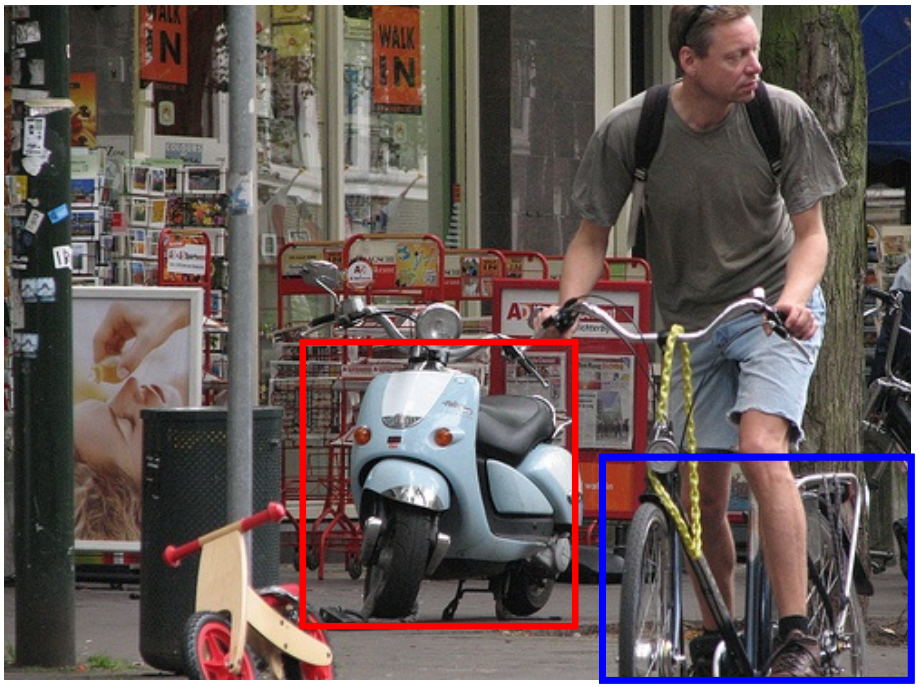
- Loss function $\Delta(y_i,\hat{y}_i)$
- Iterate

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x,\hat{y},\hat{h})}_{\text{prediction}}$$

$$\min_w \left\{ \underbrace{\Omega(w)}_{\text{regularization}} + C\sum_{i=1}^{n} \left[\underbrace{\max_{\hat{y}_i,\hat{h}_i} \left[f_w(x_i,\hat{y}_i,\hat{h}_i) + \Delta(y_i,\hat{y}_i) \right]}_{\text{most violated constraint}} - \underbrace{\max_{h_i^*} \left[f_w(x_i,y_i,h_i^*) \right]}_{\text{best latent variable}} \right] \right\}$$

- Loss function $\Delta(y_i,\hat{y}_i)$
- Iterate
  1. ...



Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x,\hat{y},\hat{h})}_{\text{prediction}}$$

$$\min_w \left\{ \underbrace{\Omega(w)}_{\text{regularization}} + C\sum_{i=1}^{n} \left[\underbrace{\max_{\hat{y}_i,\hat{h}_i} \left[f_w(x_i,\hat{y}_i,\hat{h}_i) + \Delta(y_i,\hat{y}_i) \right]}_{\text{most violated constraint}} - \underbrace{\max_{h_i^*} \left[f_w(x_i,y_i,h_i^*) \right]}_{\text{best latent variable}} \right] \right\}$$

- Loss function $\Delta(y_i,\hat{y}_i)$
- Iterate
  1. ...
  2. ...

Prediction & Parameter Learning

$$g_w(x) = \underbrace{\arg\max_{\hat{y}\in\mathcal{Y},\hat{h}\in\mathcal{H}} f_w(x,\hat{y},\hat{h})}_{\text{prediction}}$$

$$\min_w \left\{ \underbrace{\Omega(w)}_{\text{regularization}} + C\sum_{i=1}^{n} \left[\underbrace{\max_{\hat{y}_i,\hat{h}_i} \left[f_w(x_i,\hat{y}_i,\hat{h}_i) + \Delta(y_i,\hat{y}_i) \right]}_{\text{most violated constraint}} - \underbrace{\max_{h_i^*} \left[f_w(x_i,y_i,h_i^*) \right]}_{\text{best latent variable}} \right] \right\}$$

- Loss function $\Delta(y_i,\hat{y}_i)$
- Iterate
  1. ...
  2. ...
- until convergence

Experiments

Object Classification

- Graz-02, PASCAL VOC 2007, Caltech 101
- Densely extracted SIFT local features

Experiments

Object Classification

- Graz-02, PASCAL VOC 2007, Caltech 101
- Densely extracted SIFT local features

Action Classification

- Activities of Daily Living
- Harris3d detector + HOF descriptor [Laptev@CVPR08]

Experiments

Object Classification

- Graz-02, PASCAL VOC 2007, Caltech 101
- Densely extracted SIFT local features

Action Classification

- Activities of Daily Living
- Harris3d detector + HOF descriptor [Laptev@CVPR08]

		Graz-02	VOC-07	Caltech-101	Activities
Baseline	BoW	86.95 ± 1.4	49.86	61.25 ± 0.9	79.33
	SP	88.05 ± 1.4	54.74	72.68 ± 1.2	88.00
Ours	crop	88.40 ± 1.1	51.82	62.16 ± 1.0	72.00
	split	88.58 ± 1.3	55.32	73.33 ± 1.0	88.00
	crop-uni-split	90.38 ± 1.9	56.26	**75.31 ± 0.7**	**90.67**
	crop-split	**90.62 ± 1.8**	**57.05**	74.93 ± 0.9	88.67

Experiments

Object Classification

- Graz-02, PASCAL VOC 2007, Caltech 101
- Densely extracted SIFT local features

Action Classification

- Activities of Daily Living
- Harris3d detector + HOF descriptor [Laptev@CVPR08]

		Graz-02	VOC-07	Caltech-101	Activities
Baseline	BoW	86.95 ± 1.4	49.86	61.25 ± 0.9	79.33
	SP	88.05 ± 1.4	54.74	72.68 ± 1.2	88.00
Ours	crop	88.40 ± 1.1	51.82	62.16 ± 1.0	72.00
	split	88.58 ± 1.3	55.32	73.33 ± 1.0	88.00
	crop-uni-split	90.38 ± 1.9	56.26	**75.31 ± 0.7**	**90.67**
	crop-split	**90.62 ± 1.8**	**57.05**	74.93 ± 0.9	88.67

- The best results are obtained with the crop-uni-split or crop-split.

Experiments

Object Classification

- Graz-02, PASCAL VOC 2007, Caltech 101
- Densely extracted SIFT local features

Action Classification

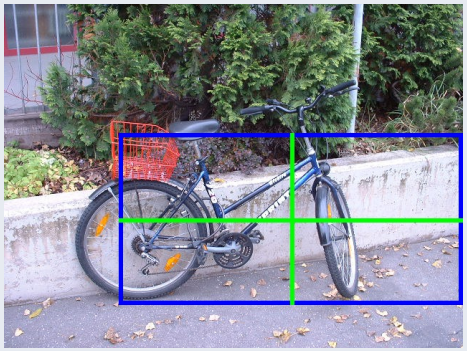
- Activities of Daily Living
- Harris3d detector + HOF descriptor [Laptev@CVPR08]

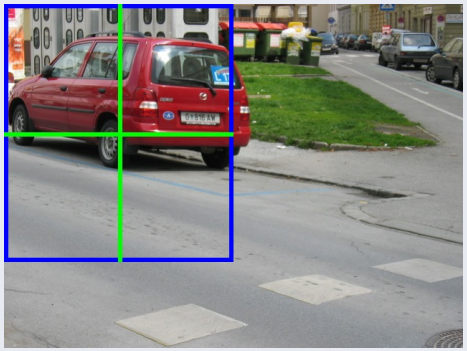
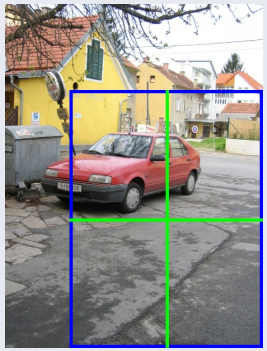
		Graz-02	VOC-07	Caltech-101	Activities
Baseline	BoW	86.95 ± 1.4	49.86	61.25 ± 0.9	79.33
	SP	88.05 ± 1.4	54.74	72.68 ± 1.2	88.00
Ours	crop	88.40 ± 1.1	51.82	62.16 ± 1.0	72.00
	split	88.58 ± 1.3	55.32	73.33 ± 1.0	88.00
	crop-uni-split	90.38 ± 1.9	56.26	**75.31 ± 0.7**	**90.67**
	crop-split	**90.62 ± 1.8**	**57.05**	74.93 ± 0.9	88.67

- The best results are obtained with the crop-uni-split or crop-split.
- Both crop-uni-split and crop-split are always better than SP.

Visualization of latent parameters

crop-uniform split examples



crop-split examples