Generalized Stacked Sequential Learning

November 13, 2014

Eloi Puertas i Prats
epuertas@ub.edu

Departament Matemàtica Aplicada i Anàlisi
Universitat de Barcelona

Universitat de Barcelona

Summary

Introduction

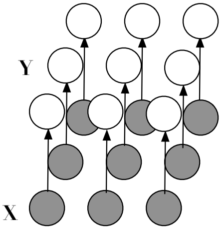
Classification: is the problem of identifying to which of a set of categories a new example belongs, given a training set of examples whose category membership is known.
Training example: (**X**,y), where **X** is a vector of features, and *y* the category which it belongs.

Classifier: $\boxed{h(\mathbf{X}) : \mathbb{R}^{\mathbf{d}} \rightarrow \mathbf{K} \in \mathbb{N}}$

Many classification problems assume that samples are **independent and identically distributed (i.i.d)**

Spam mail classification:

- Each mail is independent of each other and all of them comes from the same probability density function.



However there are problems where the i.i.d assumption does not hold. Data forms **sequences** of samples where each sample in the sequence has the same label.

- Signature classification in emails

Introduction
Sequential Learning

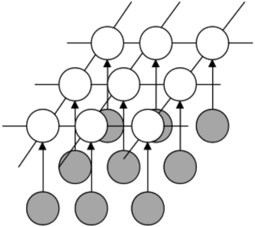
However there are problems where the i.i.d assumption does not hold. Data forms **sequences** of samples where each sample in the sequence has the same label.

- Signature classification in emails
- Laughter classification in a voice recording

However there are problems where the i.i.d assumption does not hold. Data forms **sequences** of samples where each sample in the sequence has the same label.

- Signature classification in emails
- Laughter classification in a voice recording
- **Pixel-wise classification in an image**
    - Each pixel is a sample.
    - Each sample belongs to a category.
    - Exists a spatial relationship between labels of neighboring samples.

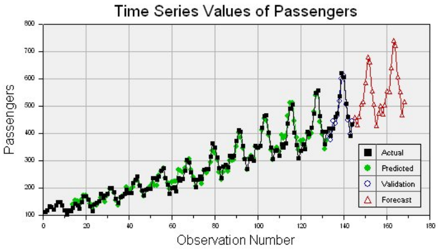
Sequential Learning: Assumes that samples are not i.i.d. Actually, pairs (**X**,y) form a sequence. Therefore it exists a kind of relationship between labels.



Sequential learning exploits the relationships between labels to improve prediction accuracy.

Sequential learning has not to be confused with:

- **Time Series:**
  - Real labels up to time t available.
  - Only need to predict label at time t+1.
  - Access to data up to time t.
- **Sequence Classification:**
  - One label expected from a full sequence
- **Non-supervised segmentation:**
  - Associated with region division according to some homogeneity criterion on the characteristics of the samples.



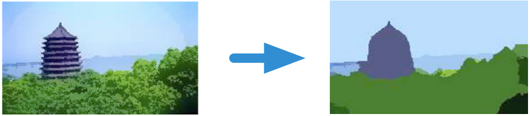
Sequential learning has not to be confused with:

- **Time Series:**
  - Real labels up to time t available.
  - Only need to predict label at time t+1.
  - Access to data up to time t.
- **Sequence Classification:**
  - One label expected from a full sequence
- **Non-supervised segmentation:**
  - Associated with region division according to some homogeneity criterion on the characteristics of the samples.

"Pagoda"

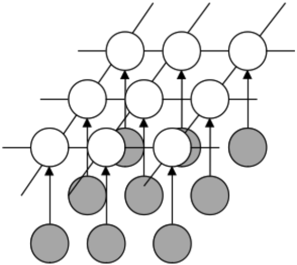
Sequential learning has not to be confused with:

- **Time Series:**
  - Real labels up to time t available.
  - Only need to predict label at time t+1.
  - Access to data up to time t.
- **Sequence Classification:**
  - One label expected from a full sequence
- **Non-supervised segmentation:**
  - Associated with region division according to some homogeneity criterion on the characteristics of the samples.



Key questions about Sequential Learning:

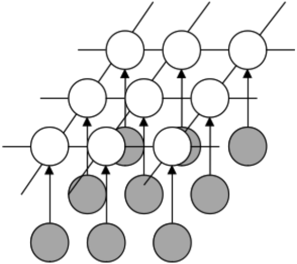
A. How to capture and exploit sequential correlations



Key questions about Sequential Learning:

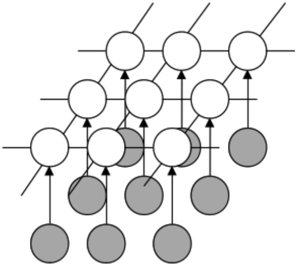
A. How to capture and exploit sequential correlations

B. How to represent and incorporate complex utility functions



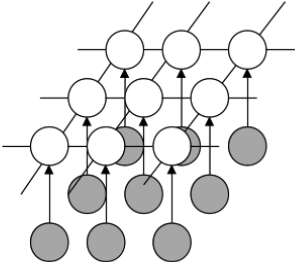
Key questions about Sequential Learning:

A. How to capture and exploit sequential correlations
B. How to represent and incorporate complex utility functions
C. How to identify long-distance interactions



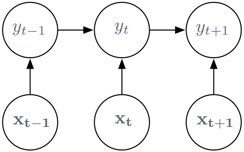
Key questions about Sequential Learning:

A. How to capture and exploit sequential correlations
B. How to represent and incorporate complex utility functions
C. How to identify long-distance interactions
D. How to make sequential learning computationally efficient



- **Markov model:** The current state depends on previous states
- Discriminative versions of Markov Models

Maximum Entropy Markov Model:

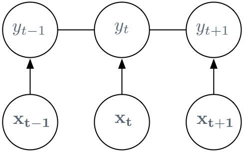


$$P(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z(\mathbf{x}, y_{t-1})} \exp\left(\sum_a \lambda_a f_a(\mathbf{x}, y_t)\right)$$

Introduction
Sequential Learning Approaches

- **Markov model:** The current state depends on previous states
- Discriminative versions of Markov Models

Conditional Random Fields

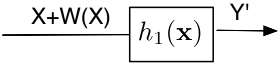


$$M_t(y_{t-1}, y_t|\mathbf{x}) = \exp(\Lambda_t(y_{t-1}, y_t|\mathbf{x}))$$

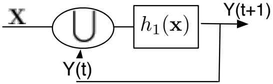
$$\Lambda_t(y_{t-1}, y_t|\mathbf{x}) = \sum_k \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}) + \sum_k \mu_k g_k(y_t, \mathbf{x})$$

$$P(y|\mathbf{x}) = \frac{\prod_{t=1}^{n+1} M_t(y_{t-1}, y_t|\mathbf{x})}{\left[\prod_{t=1}^{n+1} M_t(\mathbf{x})\right]_{\text{start,stop}}},$$

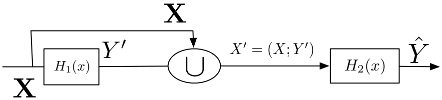
- MetaLearning strategies:
  - Sliding Windows: Maps an input window W into a single output value Y'



  - Recurrent Sliding Windows. Predicted labels are fed back to the classifier



  - Stacked Sequential Learning



Comparative:

A. How to capture and exploit sequential correlations
B. How to represent and incorporate complex utility functions
C. How to identify long-distance interactions
D. How to make sequential learning computationally efficient

Method	A	B	C	D
Maximum Entropy Markov Model	✓	—	✗	✓

Comparative:

A. How to capture and exploit sequential correlations
B. How to represent and incorporate complex utility functions
C. How to identify long-distance interactions
D. How to make sequential learning computationally efficient

Method	A	B	C	D
Maximum Entropy Markov Model	✓	—	✗	✓
Conditional Random Fields	✓	—	—	—

Comparative:

A. How to capture and exploit sequential correlations
B. How to represent and incorporate complex utility functions
C. How to identify long-distance interactions
D. How to make sequential learning computationally efficient

Method	A	B	C	D
Maximum Entropy Markov Model	✓	—	✗	✓
Conditional Random Fields	✓	—	—	—
Stacked Sequential Learning	✓	✓	—	—

$\downarrow$ neighborhood $\downarrow$ range interactions $\rightarrow$ poor performance
$\uparrow$ neighborhood $\uparrow$ long range interactions $\rightarrow$ not computationally efficient.

Proposal
Generalized Stacked Sequential Learning

Our proposal aims to achieve long range interactions, good performance and yet computationally efficient:

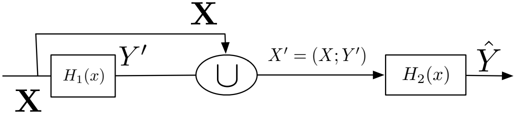
A. How to capture and exploit sequential correlations
B. How to represent and incorporate complex utility functions
C. How to identify long-distance interactions
D. How to make sequential learning computationally efficient

Method	A	B	C	D
Generalized Stacked Sequential Learning	✓	✓	✓	✓

Contributions

Generalized Stacked Sequential Learning

- A formalization of the generalized stacked sequential learning framework (GSSL).
- An extension of GSSL for the multi-class case with a compression strategy.
- An extension of GSSL specifically designed for classification of different sized objects



Settings: a window of size W, K cross-validation, base classifiers H_1, H_2

Learning algorithm: Given a trainSet X

1. $Y \leftarrow crossValidation(\mathcal{F} \leftarrow H_1(X_{N-k}), \mathcal{F}(X_k), K)$
2. $X^{ext} \leftarrow combine(X, window(Y, W))$

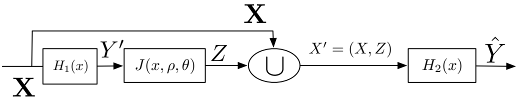
return $\mathcal{F} \leftarrow H_1(X); \mathcal{F}' \leftarrow H_2(X^{ext})$

Inference algorithm: Given an instance of TestSet **X**

1. $Y' \leftarrow \mathcal{F}(\mathbf{X})$
2. $X' \leftarrow combine(\mathbf{X}, \mathbf{window}(\mathbf{Y}', \mathbf{W}))$

return $\hat{Y} \leftarrow \mathcal{F}'(X')$

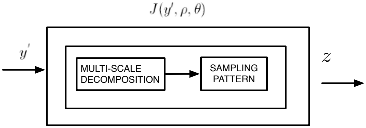
MultiScale Stacked Sequential Learning (MSSL)



A. How the relationship beetween neighbouring labels is modeled.

B. How the support lattice is created (extended set definition).

$J(Y', \rho, \theta)$ is a function that captures de data interaction with a model parameterized by θ in a neighboring ρ



the predicted labels Y' of first classifier can be defined as predicted label values or likelihood map.

A. **Multi-Scale decomposition:** each scale captures further neighborhood relationships between labels.

Multi-resolution:

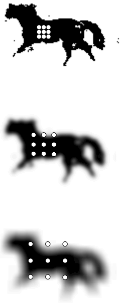
$$\Phi_{C_i}(\vec{q}; s) = y'_{C_i}(\vec{q}) * G(\sigma),$$

where G is a multidimensional Gaussian distribution with $\sigma = \gamma^s$

$\gamma = 2;$

$s \in Scales$

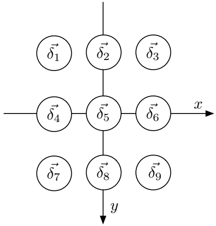
s	Multi-resolution
0	
1	
2	



A. How the support lattice is created (extended set definition).

$$z = \underbrace{\Phi(\vec{p} + \gamma^{(S-1)}\vec{\delta_1}; S), \Phi(\vec{p} + \gamma^{(S-1)}\vec{\delta_2}; S), \ldots, \Phi(\vec{p} + \gamma^{(S-1)}\vec{\delta_M}; S)}_{scale\ s=S}$$

$w = 1$



$$\delta = 3 \times \sigma$$

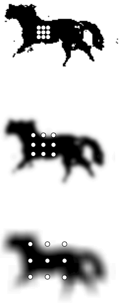
$$Size(z) = |Scales| \times (2w + 1)^d$$

s Multi-resolution

0

1

2

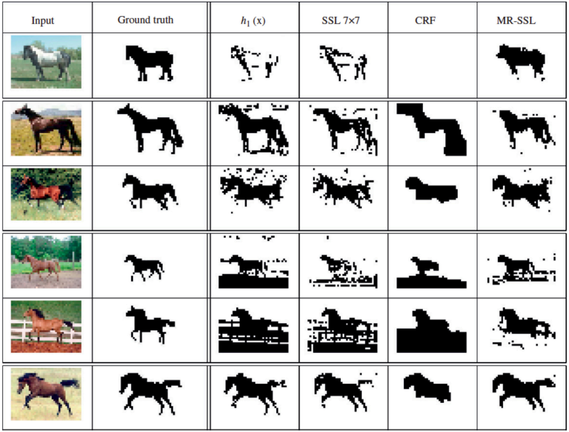


Dept. Matemàtica Aplicada i Anàlisi
Universitat de Barcelona

MultiScale SSL

Example Weizman horse DataBase

- H_1, H_2: Adaboost 100 Decision Stumps
- CIELAB components and SIFT features
- MultiScale: 7 Scales, 8 neighbours.



MultiScale SSL

Results Weizman horse DataBase

- H_1, H_2: Adaboost 100 Decision Stumps
- CIELAB components and SIFT features
- MultiScale: 7 Scales, 8 neighbours.

	Accuracy	Overlapping o
AdaBoost	$0.55(\pm 0.2398)$	$0.61(\pm 0.2064)$
CRF	$0.52(\pm 0.3257)$	$0.5137(\pm 0.2641)$
SSL 7×7	$0.63(\pm 0.2308)$	$0.56(\pm 0.2264)$
MR-SSL	**$0.8592(\pm 0.0903)$**	**$0.6819(\pm 0.2109)$**

MultiClass MSSL

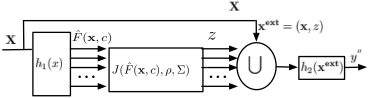
MultiClass classification problems

Generally classification problems are not just binary.

MultiClass classifiers, besides a prediction, it can provides one likelihood value for each class:

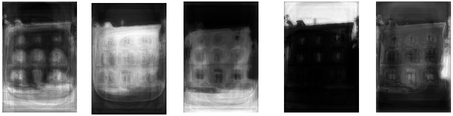
1 Other 2 Building 3 Sky 4 Road 5 Vegetation

MultiClass MSSL

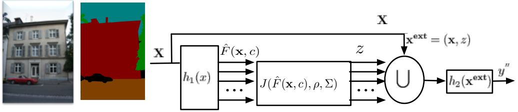


- Likelihood map for each class

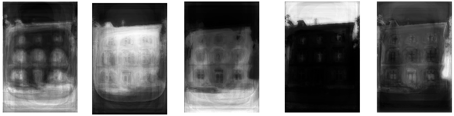
- Multiscale decomposition for each likelihood map



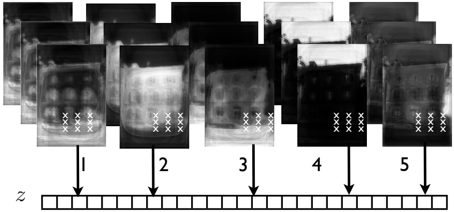
MultiClass MSSL



- Multiscale decomposition for each likelihood map



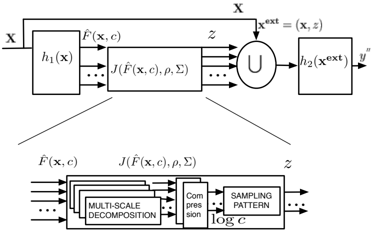
- Sampling and vector of extended features



size: $(2w+1)^d \times |\Sigma| \times |\mathcal{L}|$

Drawback: number of extended features increases with number of classes.

- Compression strategy:



Reduce $|z|$:

$$|z| = |\mathcal{L}| \times |\Sigma| \times (2w+1)^d \rightarrow |z| = \lceil \log_2 |\mathcal{L}| \rceil, |\Sigma| \times (2w+1)^d$$

- Build a discriminative code that groups $\mathcal{L}$ in different partitions

MultiClass MSSL

Compression

$$\hat{F}^{s_k}(\mathbf{x}, P_j) = \sum_i^N \gamma_{ij} \hat{F}^{s_k}(\mathbf{x}, c_i), \qquad \gamma_{ij} = \begin{cases} 0 & -1 & \text{if } c_i \in P_j^1 \\ 1 & 1 & \text{if } c_i \in P_j^2 \end{cases}$$

Binary compression

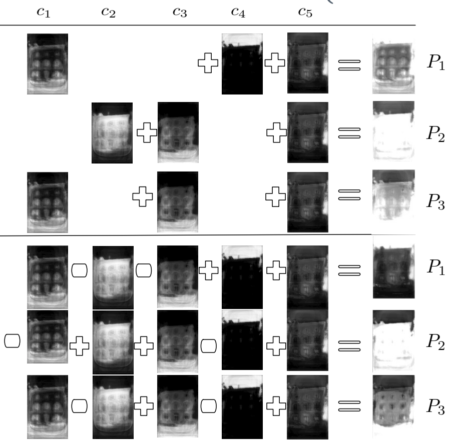
	c_1	c_2	c_3	c_4	c_5
Γ_1	1	0	0	1	1
Γ_2	0	1	1	0	1
Γ_3	1	0	1	0	1

Table 1.

Ternary compression

	c_1	c_2	c_3	c_4	c_5
Γ_1	1	-1	-1	1	1
Γ_2	-1	1	1	-1	1
Γ_3	1	-1	1	-1	1

Table 2.



Experiments
Settings

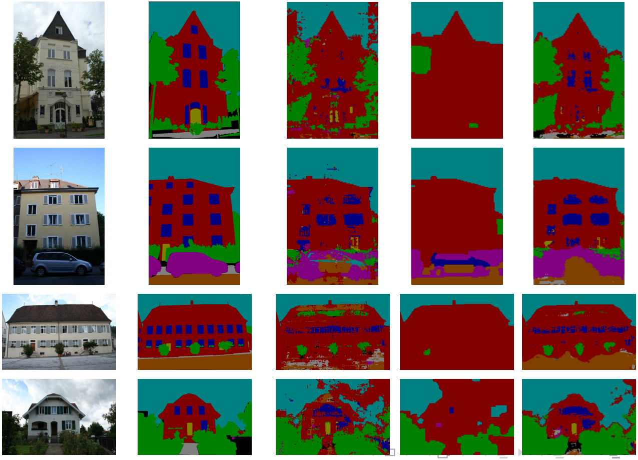
- DataSets:
  - eTrims 4-Classes. RGB and RGB+HOG 9 bins features.
  - eTrims 8-Classes. RGB and RGB+HOG 9 bins features.
  - IVUS image: 8-Classes 29 textural features.
  - FAQ dataset: 4-Classes 1D. 24 boolean features.
- H_1 and H_2 Real Adaboost 100 decision Stumps + ECOC *one-versus-one*
- 4 Scales
- Multi-label optimization via α-expansion from H_1 confidences map.

Experiments

Etrims 8- Classes

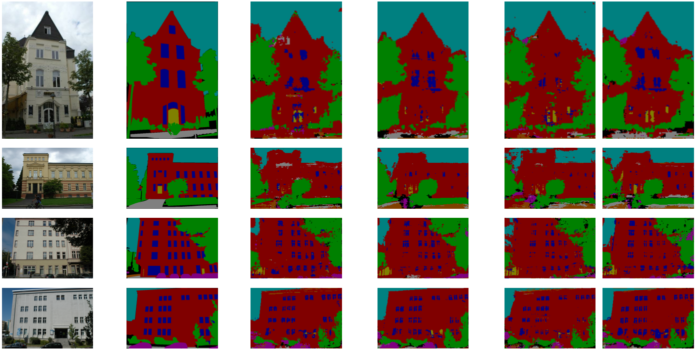
ETRIMS 8 Classes HOG database

Original Groundtruth ADAboost GraphCut MMSSL



Comparative between MMSSL approaches in ETRIMS 8 Classes HOG database.

Original Groundtruth Labels Confidences Binary Ternary



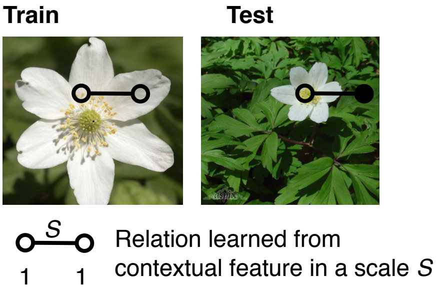
Experiments
Etrims 8- Classes

Result figures for database ETRIMS 8 classes RGB and HOG.

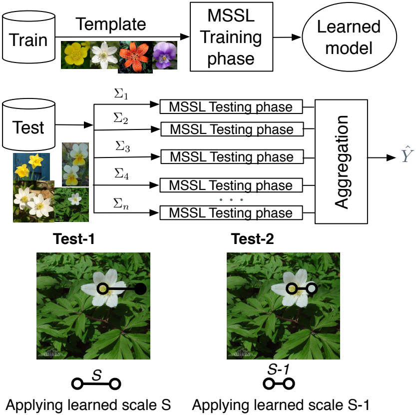
		Accuracy	Overlapping
RGB	ADABoost	0.606	0.1991
	GraphCuts	0.6039	0.1859
	Labels	0.6549	0.2526
	Standard	**0.703**	**0.3133**
	SublinealBinary	0.6616	0.267
	SublinealTernary	0.6742	0.2768
HOG	ADABoost	0.6723	0.2868
	GraphCuts	0.6812	0.2618
	Labels	0.6885	0.3031
	Standard	**0.7312**	**0.3479**
	SublinealBinary	0.6895	0.3038
	SublinealTernary	0.7164	0.3348

MSSL learns relationships between pairs of example and labels, making them explicit.

- **Problem:** relationships between pairs changes if sequences sizes changes.
- **Example:** Pixel-wise flower classification:



Sequential Learning at different sizes
Shifting approach



Train:

- Use templates (images of the same size).
- Choose a set of L consecutive scales $\Sigma_T \in \Sigma = \{\sigma = 0, \ldots, \sigma = N\}$

Test:

- Perform several testing phases shifting scales.
- Choose the $N - L$ set of L consecutive scales $\Sigma_1 \ldots \Sigma_n \in \Sigma$ shifting one position.

Aggregation:

- Maximum likelihood value for each pixel

Experiments

Flower dataset

Settings:

- Train Scales: $\sigma \in \{18, 27, 41\}$
- Test Scales: $\sigma \in \{0.5, 3, 5, 8, 12, 18, 27, 41\}$.
- Test Rounds: 6
  $\sigma \in \{0.5, 3, 5\}, \sigma \in \{3, 5, 8\} \ldots \sigma \in \{18, 27, 41\}$

Results using Adaboost and MSSL.

Method	Acc	Over
ADABoost	0,8773	0,5621
CRF	0,8568	0,5840
Shift MSSL	**0,9012**	**0,6243**

Generalized Stacked Sequential Learning

Eloi Puertas

Introduction

Generalized SSL

MultiScale SSL

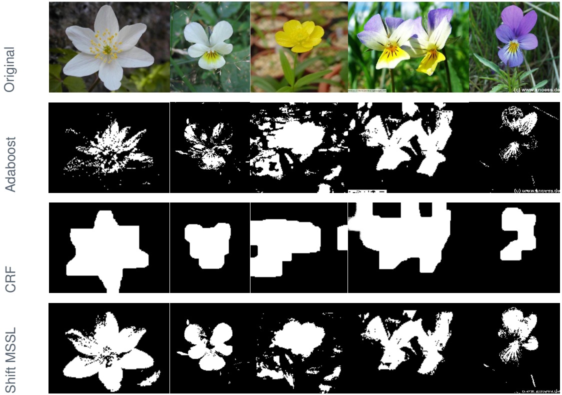
MultiClass MSSL

Shifting

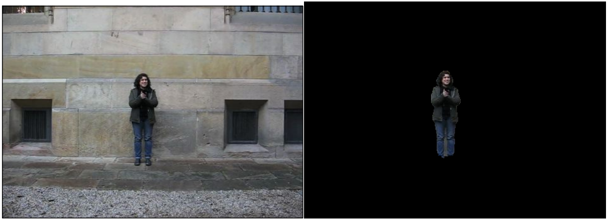
Application

Conclusions

Predictions



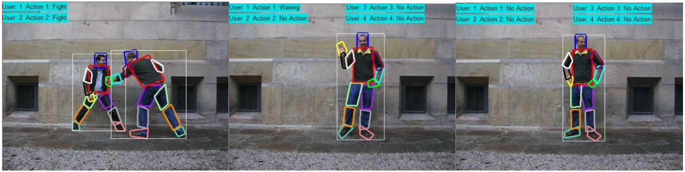
- **Problem:** Segmenting the human body in still RGB images.
- Several people can appear portraying a wide range of poses, lighting conditions,clothes...



Approach
Human body segmentation

A way to approach this problem in literature:

- First stage (body-part detection):
  - Use base classifier to learn body parts: SVM, Adaboost, Cascading Classifiers.
  - A noisy set of candidate parts is obtained.
- Second stage (joint optimization with body constraints)
  - Probabilistic Graphical Models are used to find the most probable pose (PS, Poslets, CRF).
  - Typically this approaches yield a multi-limb detection of the pose.



Contribution
Human body segmentation

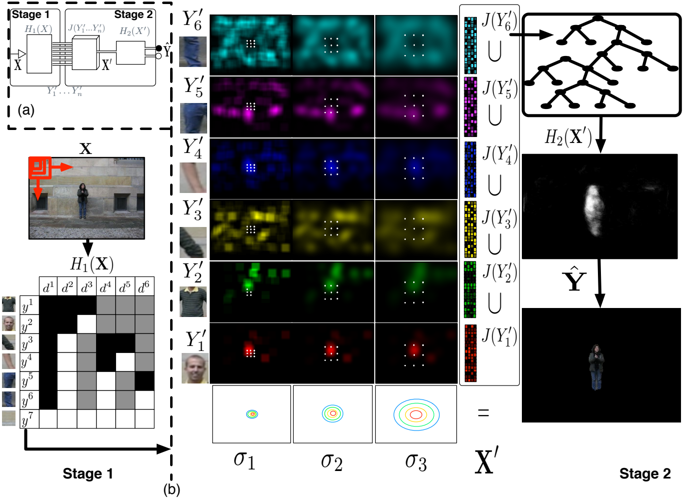
Our contribution:

- Improve the **binary** segmentation of the human body in RGB images by learning context-aware features.

Using the same two stage scheme:

- First stage uses a base classifier to learn body parts
- Use the MSSL framework to build the extended feature set
- Obtain a prior pixel-wise binary classification of the image (person vs. background)

Pipeline
Human body segmentation

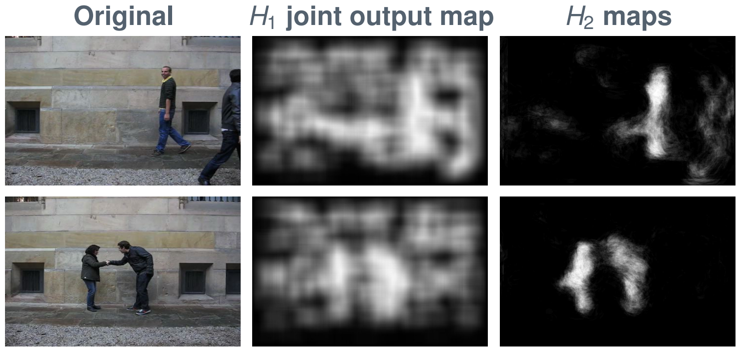


Settings:

- Dataset: HuPBA 8k+, more than 8000 images with 14 human body parts manually labeled.
- H_1 = Cascade of Classifiers based on AdaBoost and Haar-like features.
- Scales = 3 with $\sigma = [8, 16, 32]$ for each body part (label)
- H_2 = Random Forest with 50 decision trees.

	MSSL	H_1 **Soft Detector**
Overlapping	**62,73**	59,33

Experiments
Human body segmentation



Conclusions

This thesis focuses on the problem of sequential learning from a meta-learning perspective.

Contributions

- A generalization of the stacked sequential learning framework (SSL) stressing the key role of the neighborhood modeling
- A general extension of MSSL for the multi-class case with a compression strategy
- An extension of MSSL specifically designed for classification of different sized objects
- An application of MSSL for human body segmentation

Future Work

- To use other compression approaches, such as compressed sensing or PCA, for the sake of reducing the extended set as much as possible without losing accuracy.
- To consider a scale and rotation invariant architecture, shifting not only the scales but also the sampled neighborhood patterns.
- To refine our body segmentation application in order to perform accurate multi-limb body segmentation. This is, to perform correctly segmentation of each part of the body (arms, legs, torso, head...) by using our MSSL framework.
- To analyse the improvement obtained by using several levels of stacking in GSSL
- To find a theoretical correspondence between GSSL and other related sequential learning frameworks such convolutional neural network.

Publications I

- (2009) Multi-modal laughter recognition in video conversations. S Escalera, E Puertas, P Radeva, O Pujol. *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops. IEEE Computer Society Conference on,*
- (2009) Multi-scale stacked sequential learning. O Pujol, E Puertas, C Gatta. *Multiple Classifier Systems*, 262-27.
- (2009) Multi-Scale Multi-Resolution Stacked Sequential Learning. E Puertas, C Gatta, O Pujol. *Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence (CCIA)*. 112-117.
- (2010) Classifying Objects at Different Sizes with Multi-Scale Stacked Sequential Learning. E Puertas, S Escalera, O Pujol *Proceedings of the 10th International Conference of the Catalan Association for Artificial Intelligence (CCIA)*, 193-200.

Publications II

- (2011) Multi-scale stacked sequential learning. C Gatta, E Puertas, O Pujol. *Pattern Recognition* 44 (10), 2414-2426
- (2011) Multi-class multi-scale stacked sequential learning. E Puertas, S Escalera, O Pujol. *Multiple Classifier Systems*, 197-206
- (2013) Generalized multi-scale stacked sequential learning for multi-class classification. E Puertas, S Escalera, O Pujol. *Pattern Analysis and Applications*, 1-15
- (2014) Learning to segment humans by stacking their body parts. E Puertas, M.A Bautista, D S'anchez, S Escalera, O Pujol. *ChaLearn Looking at people Workshop, ECCV*

Thank you for your attention!

Universitat de Barcelona