Weakly supervised discriminative localization and classification: a joint learning process

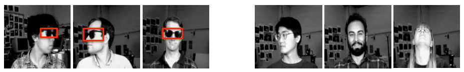
Minh Hoai Nguyen[1], Lorenzo Torresani[2], Fernando de la Torre[1], & Carsten Rother[3]

[1]Carnegie Mellon University, [2]Dartmouth College, [3]Microsoft Research Cambridge

Motivation

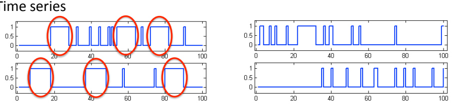
Classification problems

- Glasses vs. no-glasses

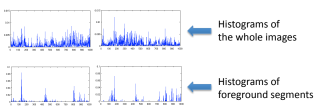


- Cars vs. no-car

- Time series

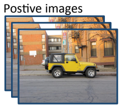


Global vs. local statistics

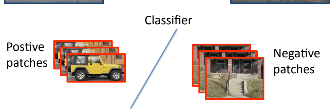


Traditional solution: classification based on subregions

Postive images

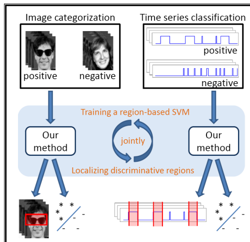


Negative images



- Require ground truth manual selections:
  - Time consuming
  - Inconsistent
  - Non-optimal

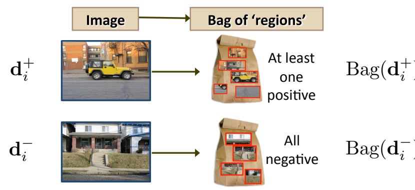
Overview

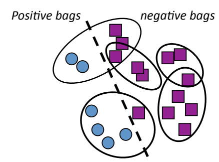


- Joint localization & classification
- Training requires class labels but ***not*** region selections

Proposed method

Multiple Instance Learning:





Maron-Ratan-ICML98, Yang-Lozano-Perez-ICDE00, Andrews-et-al-NIPS03, Chen-Wang-JMLR04

Localization-classification SVM: the formulation

$$\underset{\mathbf{w},b}{\text{minimize}} \quad \frac{1}{2}||\mathbf{w}||^2$$

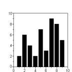
$$\text{s.t.} \quad \max_{\mathbf{x}\in \mathrm{Bag}(\mathbf{d}_i^+)} \mathbf{w}^T\varphi(\mathbf{x}) + b \geq 1 \;\forall i$$

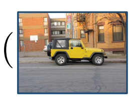
$$\max_{\mathbf{x}\in \mathrm{Bag}(\mathbf{d}_i^-)} \mathbf{w}^T\varphi(\mathbf{x}) + b \leq -1 \;\forall i$$

Optimization

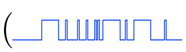
- Two iterative loops
- Outer loop – coordinate descent: alternate between optimizing (**w**, b) and the instances of positive bags that maximize the SVM scores.
- Inner loop – constraint generation: add the most violated constraints into the constraint sets.
- Each iteration requires $\hat{\mathbf{x}} = \underset{\mathbf{x}\in \mathrm{Bag}(\mathbf{d})}{\operatorname{argmax}}\, \mathbf{w}^T\varphi(\mathbf{x})$

Representation & Efficient localization

$\varphi(\mathbf{x}) =$  Histogram of visual words

$\mathrm{Bag}($  $) =$ all possible subwindows

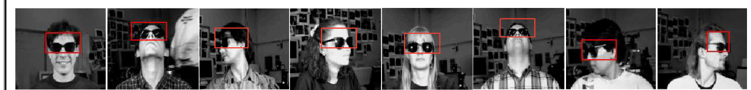
Efficient search using branch-and-bound
Average 100ms/image (480*640 pixel)

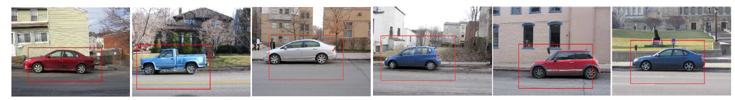
$\mathrm{Bag}($  $) = \{$ At most ***k*** disjoint intervals $\}$

Efficient search using dynamic programming
Average 10ms/signal (15000 frames)

Experiments

Images – Localization results





Failure cases

Images – Quantitative results

Dataset	Measure	Bag of words	SVM with global statistics	SVM with human labels	Ours
Faces	Acc. (%)	80.11	82.97	86.79	**90.0**
	ROC Area	n/a	0.90	0.94	**0.96**
Cars	Acc. (%)	77.5	80.75	81.44	**84.0**
	ROC Area	n/a	0.86	0.88	**0.90**
Airplanes	Acc. (%)	89.74	**96.05**	89.40	**96.05**
	ROC Area	n/a	**0.99**	0.95	**0.99**
Cars	Acc. (%)	94.93	98.17	n/a	**98.28**
	ROC Area	n/a	**1.00**	n/a	**1.00**
Faces	Acc. (%)	59.83	88.70	86.78	**89.57**
	ROC Area	n/a	**0.95**	0.91	**0.95**
Motorbikes	Acc. (%)	76.80	**88.99**	84.67	87.81
	ROC Area	n/a	**0.95**	0.92	0.94

Caltech-4 datasets (Airplanes, Cars, Faces, Motorbikes)

Several Conclusions

- Human labels often are not optimal
- Tight bounding boxes often are not optimal; contextual information is important.
- Segmentation does not always help. Our method determines automatically the optimal support region for classification

Synthetic time series data

max # of disjoint intervals allowed

Using global statistics: ROC: 0.577, Acc: 66.5%

k	1	2	3 to 7	8	12	16	20
Acc.(%)	77.0	93.0	100	98.5	91.5	77.5	67.25
ROC Area	.843	.980	1.00	.998	.933	.793	.613

Several Conclusions

- Segmentation does help.
- Multiple disjoint intervals are necessary.
- Classification performance is not too sensitive to the number of maximum disjoint intervals allowed.

Mouse activity classification

Action	Dollár et al. [8]	1-NN	SVM	Ours
Drink	0.63	0.58	0.63	**0.67**
Eat	**0.92**	0.87	0.91	0.91
Explore	0.80	0.79	**0.85**	**0.85**
Groom	0.37	0.23	0.44	**0.54**
Sleep	0.88	0.95	**0.99**	**0.99**

F_1 score

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$