WaldBoost: Learning for Sequential Classification

Jan Šochman

joint work with Jiří Matas

Center for Machine Perception
Czech Technical University, Prague
http://cmp.felk.cvut.cz

Motivation

We know that...

- Time-to-decision vs. precision trade-off is inherent in many computer vision problems
- Decision time clearly influences impact of a method:
  - Viola-Jones (2001) – real-time performance 2500 citations
  - Schneiderman-Kanade (1998) – smaller error rates, but 1000x slower 250 citations

But...

- Machine learning focuses mainly on error minimisation
- Time-to-decision vs. precision trade-off not explicitly stated in the problem formulation
- Speedup usually gained heuristically by code optimisation, hardware implementation, new architecture proposals, ...
- A sequential decision theory is well established in statistics, yet it is not formulated as a *learning task*

Talk outline

- Learning problem formulation
- WaldBoost algorithm
- Applications
  - Fast face detection
  - Interest point detector speedup by emulation

Sequential Classification

Basic notions

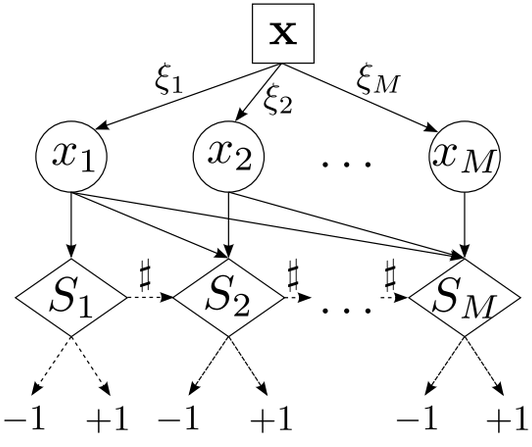
$\mathbf{x} \in \mathcal{X}$ — classified object (e.g. image)

$y \in \mathcal{Y} = \{-1, +1\}$ — object class

$(\xi_1, \ldots, \xi_M)$ where $\xi_t \colon \mathcal{X} \to \mathcal{X}_t$ — vector of measurement functions

$(x_1, \ldots, x_M)$ where $x_t \in \mathcal{X}_t$ — ordered sequence of measurements

$S = (S_1, \ldots, S_M)$ where $S_t : \mathcal{X}_1 \times \ldots \times \mathcal{X}_t \to \mathcal{Y} \cup \{\sharp\}$ — sequential strategy



Any sequential strategy S is associated with

- $\alpha_S = P(S = -1 | y = +1), \qquad \beta_S = P(S = +1 | y = -1)$ — error of the first and the second kind
- $\bar{T}_S^+ = E_{\mathcal{X}^+}[T_S(\mathbf{x})] \qquad \bar{T}_S^- = E_{\mathcal{X}^-}[T_S(\mathbf{x})]$ — average evaluation time for each class

  where $T_S(\mathbf{x}) = \arg\min_t (S_t(\xi_1(\mathbf{x}), \ldots, \xi_t(\mathbf{x})) \neq \sharp)$ — time-to-decision

Classification task

For given bounds α and β, find a strategy S^* such that

$$\alpha_{S^*} \leq \alpha \text{ and } \beta_{S^*} \leq \beta$$

and for any strategy S

$$\bar{T}_{S^*}^+ \leq \bar{T}_S^+ \text{ and } \bar{T}_{S^*}^- \leq \bar{T}_S^- \qquad \text{s.t. } \alpha_S \leq \alpha \text{ and } \beta_S \leq \beta$$

Optimal Sequential Test

Sequential Probability Ratio Test (SPRT) [Wald 1947]

SPRT is a sequential strategy S^*

$$S_t^* = \begin{cases} +1, & R_t \leq B \\ -1, & R_t \geq A \\ \sharp, & B < R_t < A \end{cases} \qquad R_t = \frac{p(x_1, ..., x_t | y = -1)}{p(x_1, ..., x_t | y = +1)}.$$

Practical (and nearly optimal) setting for the thresholds A and B

$$A = \frac{1-\beta}{\alpha}, \qquad B = \frac{\beta}{1-\alpha}$$

Difficulties

SPRT is an optimal sequential test in sense of our optimisation problem **but**:

1. the ordering of the measurements is expected to be given
2. multidimensional pdf's $p(x_1, ..., x_t | y = \textit{class})$ has to be computed in non i.i.d case

Learning Problem Formulation

Given

- a training set $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ $\quad \mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, +1\}$
- an unordered set (class) of measurement functions
$$\Xi_u = \{\xi_1, \xi_2, \ldots, \xi_M\} \qquad \xi_t \colon \mathcal{X} \to \mathcal{X}_t$$
- two values α, β such that $0 \leq \alpha, \beta \leq 1$

Find

- an ordering π of Ξ_u
- a sequential strategy S^L using $(\xi_{\pi(1)}, \ldots, \xi_{\pi(M)})$

such that

$\alpha_{S^L} \leq \alpha$ and $\beta_{S^L} \leq \beta$

and for any sequential strategy S using Ξ_u

$\bar{T}^+_{S^L} \leq \bar{T}^+_S$ and $\bar{T}^-_{S^L} \leq \bar{T}^-_S$ $\qquad$ s.t. $\alpha_S \leq \alpha$ and $\beta_S \leq \beta$

How to keep the optimality of SPRT?

- ordering: What is a good measurement ordering?
- decision thresholds: How to avoid multi-dimensional density estimation?

WaldBoost: Measurement Selection and Ordering

Why to select and order the measurements by AdaBoost

- AdaBoost can be seen as a feature (measurement) selector with a principled strategy (minimisation of upper bound on empirical error)
- Is close to sequential decision making, as it produces a sequence of gradually more complex classifiers
- Its output converges to the logarithm of likelihood ratio

Real AdaBoost [Schapire & Singer, ML 1999]

- Selects iteratively weak classifiers $h_t \colon \mathcal{X} \to \mathbb{R}$ and combines their outputs into a strong classifier

$$f_T(\mathbf{x}) = \sum_{t=1}^{T} h_t(\mathbf{x})$$

- Classification is given by $sign(f_T(\mathbf{x}))$

Measurements

Each weak classifier is associated with one measurement $\xi_t \equiv h_t$ $\qquad \mathcal{X}_t = \mathbb{R}$

$\Rightarrow h_1, h_2, \ldots$ **ordered sequence of measurements**

WaldBoost: Learning the Decision Thresholds

Ratio needed for SPRT

$$R_t(\mathbf{x}) = \frac{p(h_1(\mathbf{x}), ..., h_t(\mathbf{x})|y=-1)}{p(h_1(\mathbf{x}), ..., h_t(\mathbf{x})|y=+1)}$$

- Weak classifiers not independent $\rightarrow$ multi-dimensional $\rightarrow$ intractable
- We would like to have a test working with $f_t(\mathbf{x})$ directly

Asymptotic convergence of AdaBoost [Friedman, TR 1998]

$$\lim_{t\to\infty} f_t(\mathbf{x}) = -\frac{1}{2}\log R(\mathbf{x}) + \frac{1}{2}\log\frac{P(y=+1)}{P(y=-1)}$$

$R(x) = \frac{p(\mathbf{x}|y=-1)}{p(\mathbf{x}|y=+1)}$

- AdaBoost response $f_t(\mathbf{x})$ is (asymptotically) a motonic function of $R(\mathbf{x})$
- Thresholding $R_t(\mathbf{x})$ on $A \Leftrightarrow$ thresholding $f_t(\mathbf{x})$ on some θ_t^A (similarly for B and θ_t^B)

for sufficiently large t

- Knowing the thresholds θ_t^A and θ_t^B, the sequential decision strategy would become

$$S_t = \begin{cases} +1, & f_t(\mathbf{x}) \geq \theta_t^B \\ -1, & f_t(\mathbf{x}) \leq \theta_t^A \\ \sharp, & \theta_t^A < f_t(\mathbf{x}) < \theta_t^B \end{cases}$$

inequalities inverted
because $f(\mathbf{x}) \approx -R(x)$

WaldBoost: Learning the Decision Thresholds

Search procedure (for θ_t^A only)

- We want to find θ_t^A corresponding to the condition

$$R_t(\mathbf{x}) = \frac{p(h_1(\mathbf{x}), ..., h_t(\mathbf{x})|y=-1)}{p(h_1(\mathbf{x}), ..., h_t(\mathbf{x})|y=+1)} \geq A$$

classification to -1 class

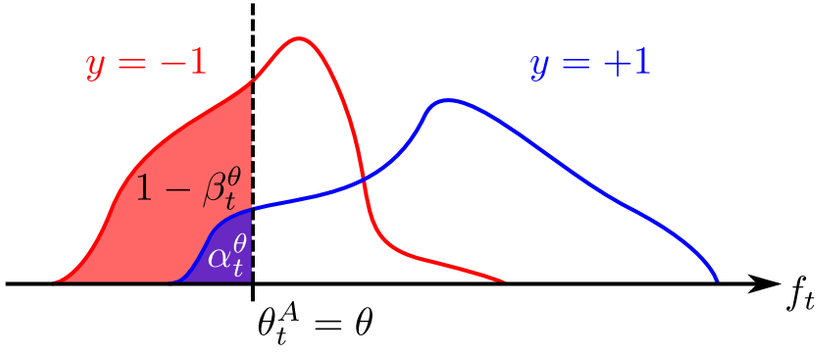
- This can be rewritten as

$$\underbrace{\sum_{\mathbf{x}\in D_t^-} p(h_1, \ldots, h_t|y=-1)}_{1-\beta_t \text{ (true negatives)}} \geq A \underbrace{\sum_{\mathbf{x}\in D_t^-} p(h_1, \ldots, h_t|y=+1)}_{\alpha_t \text{ (false negatives)}}$$

$D_t^- = \{\mathbf{x}; R_t(\mathbf{x}) \geq A\}$

- Using the asymptotic property

$$D_t^- = \{\mathbf{x}; R_t(\mathbf{x}) \geq A\} = \{\mathbf{x}; f_t(\mathbf{x}) \leq \theta_t^A\}$$



Search for θ such that

$$1 - \beta_t^\theta \geq A\alpha_t^\theta$$

WaldBoost Learning

Input:

- A training set $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ $\quad \mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, +1\}$
- A set (class) of weak classifiers $\mathcal{H} = \{h_i\}_{i=1}^{M}$ each one associated to one measurement
- Desired final false negative rate α and false positive rate β
- The number of iterations T

Set $A = (1 - \beta)/\alpha$ and $B = \beta/(1 - \alpha)$

for $t = 1, ..., T$

1. Find $h_t \in \mathcal{H}$ by AdaBoost using $\mathcal{T}$ and add it to the strong classifier $f_t = f_{t-1} + h_t$
2. Find the decision thresholds θ_t^A and θ_t^B for f_t
3. Remove from training samples $\mathbf{x} \in \mathcal{T}$ for which $f_t(\mathbf{x}) \geq \theta_t^B$ or $f_t(\mathbf{x}) \leq \theta_t^A$
4. Sample new data into the training set $\mathcal{T}$

end

Output:

- Ordered set of weak classifiers $\{h_t\}_{t=1}^{T}$
- The decision thresholds $\{\theta_t^A\}_{t=1}^{T}$ and $\{\theta_t^B\}_{t=1}^{T}$

WaldBoost Classification

Given: h_t, θ_t^A, θ_t^B, $\gamma \quad (t = 1, \ldots, T)$

Input: an object $\mathbf{x}$

For $t = 1, \ldots, T$ (SPRT execution)

If $f_t(\mathbf{x}) \geq \theta_t^B$, classify $\mathbf{x}$ to the class $+1$ and terminate

If $f_t(\mathbf{x}) \leq \theta_t^A$, classify $\mathbf{x}$ to the class -1 and terminate

end

If $f_T(\mathbf{x}) > \gamma$, classify $\mathbf{x}$ as $+1$. Classify $\mathbf{x}$ as -1 otherwise.

Non-Symmetric WaldBoost Decisions

Non-symmetric decision problems (e.g. face detection)

- Positive (face) and negative (background) class complexities are unbalanced
- Positive samples difficult to collect
- Positive samples appear rarely during classification
- Missed detection rate more serious than false positive rate

Set $\beta = 0$ in WaldBoost learning. Then

SPRT thresholds

$$A = \frac{1-0}{\alpha} = \frac{1}{\alpha}$$
$$B = \frac{0}{1-\alpha} = 0$$

SPRT strategy

$$S_t^* = \begin{cases} +1, & R_t \leq 0 \\ -1, & R_t \geq 1/\alpha \\ \sharp, & 0 < R_t < 1/\alpha \end{cases}$$

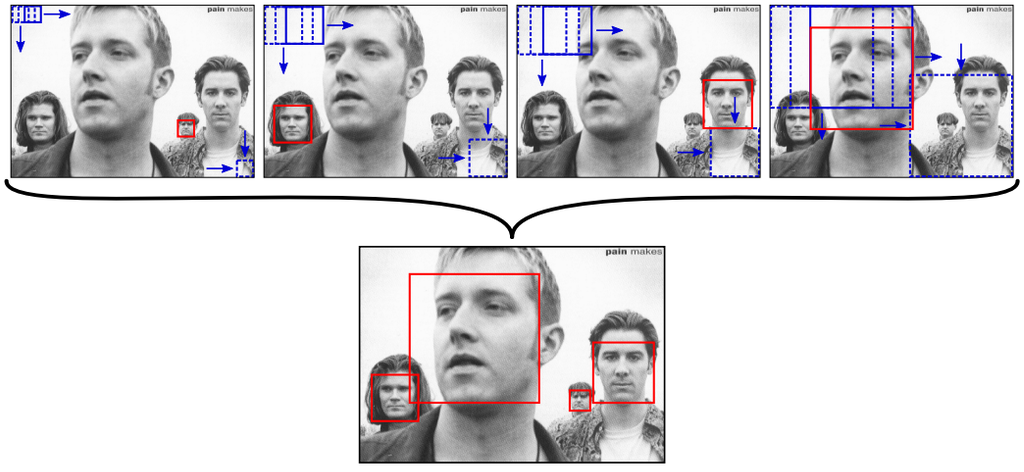
WaldBoost strategy

$$S_t = \begin{cases} -1, & f_t(\mathbf{x}) \leq \theta_t^A \\ \sharp, & \theta_t^A < f_t(\mathbf{x}) \end{cases}$$

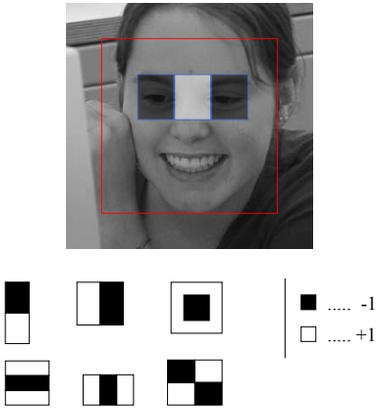
- Since $R_t(\mathbf{x}) > 0$, only decisions to the class -1 (background) allowed
- Only the -1 class pruned during trainig
- All measurements evaluated for positive samples (does not affect the speed much since they are rare)

Application 1: Fast Face Detection

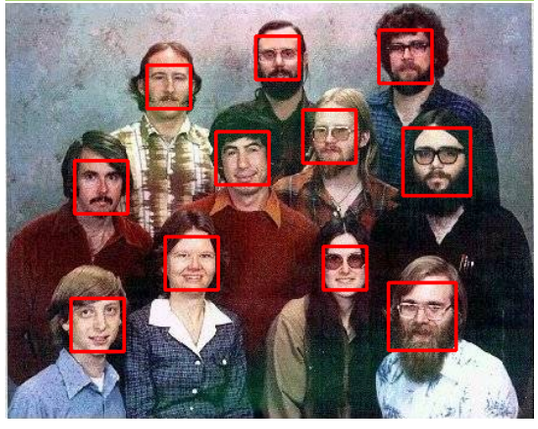
scanning process



used measurements



results

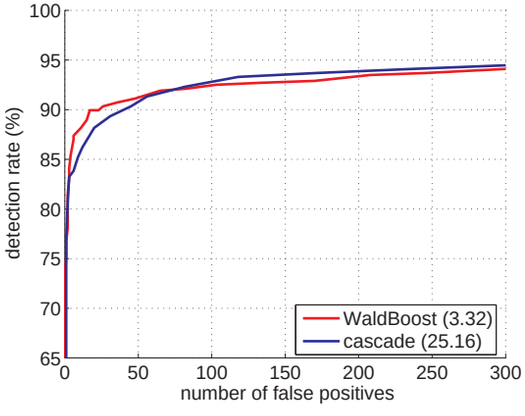


Results

- State-of-the-art detection results
- Fastest detector among those with comparable detection rate
- Principle way of learning a cascade-like classifier

  only two global parameters: α and β
- Web demo

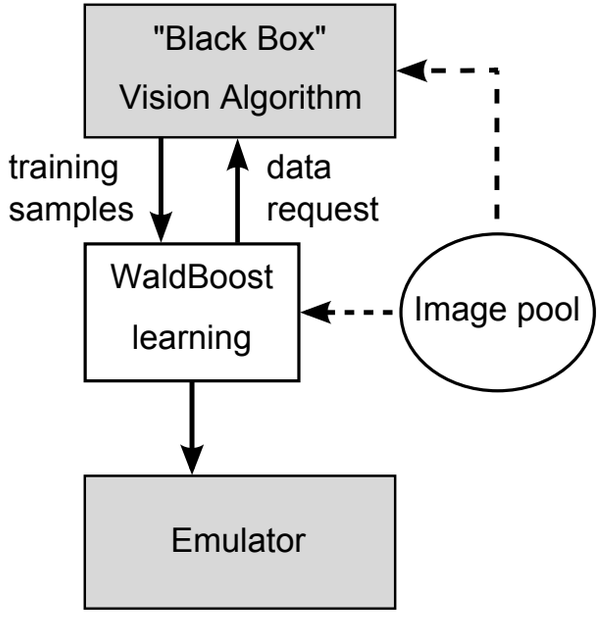
http://cmp.felk.cvut.cz/demos/FaceDetection/



Application 2: Fast Emulators of Interest Point Detectors

Idea

- Assume: slow binary-valued "black-box" algorithm is available
- Use the black-box algorithm to generate a training set
- WaldBoost optimises:
  - output similarity
  - evaluation speed

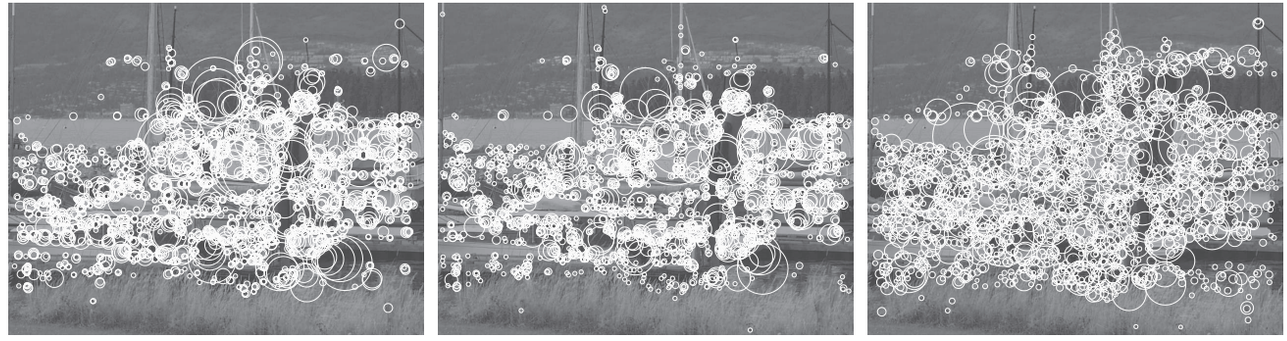


Results

- A general framework for speeding up existing algorithms by a sequential classifier learned by the WaldBoost algorithm
- Demonstrated on Hessian-Laplace and Kadir-Brady saliency detectors

	HL	KB
original	0.9s	1m 48s
SURF	0.09s	—
speed-up	10×	—
WaldBoost	0.10s	0.76s
speed-up	9×	142×

Hessian-Laplace WaldBoost SURF



Conclusions

- Time to decision vs. precison trade-off problem was formalised as a constrained optimisation problem
- WaldBoost overcomes limitations of SPRT to a priori ordered measurements and known multi-dimensional pdf's
- Even though WaldBoost is greedy (sub-optimal), it is highly flexible and has interesting applications
  - Fast face detection
  - Fast emulators of interest point detectors
- Other related papers
  - Randomized RANSAC with SPRT [Matas & Chum, ICCV 2005]
  - On-line WaldBoost for tracking [Grabner et al., ICPR 2008]
  - Sequential correspondence selection by cosegmentation [Čech et al., PAMI 2009]

Thank you for attention!

The End