Multi-Stage Decision System

Kirill Trapeznikov
(Venkatesh Saligrama and David Castañón, Boston University)

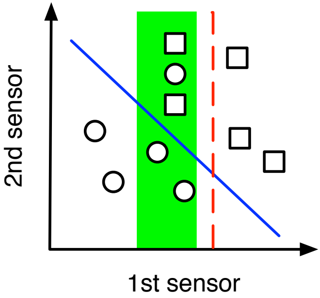
October 24th, 2012

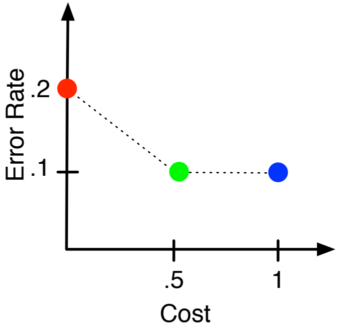
Overview

- Objective: reduce measurement cost in decision systems without performance degradation by using adaptive sensing
  - Adaptively collect measurements from different sensors based on collected observations
  - Not all decisions require every sensor measurement
  - Reduce average sensing cost to meet budget
- Result: Novel Multi-Stage Classifier Design Framework
  - A non-parametric theory for training adaptive classification systems directly from data
  - Extends existing Machine Learning (ML) techniques
  - Suitable for both detection and multi-class decisions
- Illustrate performance with experiments on collected data
  - Datasets from UCI ML Repository
  - Concealed explosive detection data (Courtesy of SAIC, S. Macintosh)
  - Results show optimal performance with reduced budgets, superior to that of alternative adaptive classifier designs

Are all sensors necessary to classify every sample?

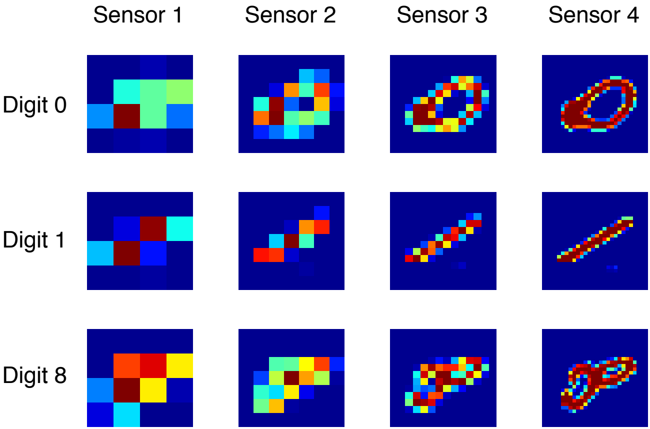
Some samples can be classified using only low cost sensor





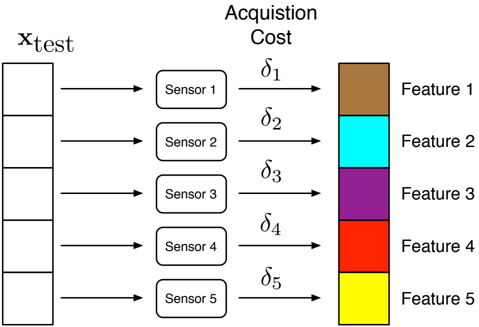
Sensor	Cost
1	0
2	1

Strategy needs to be adaptive



Sensor requirement is sample dependent

Sensors have different acquisition costs



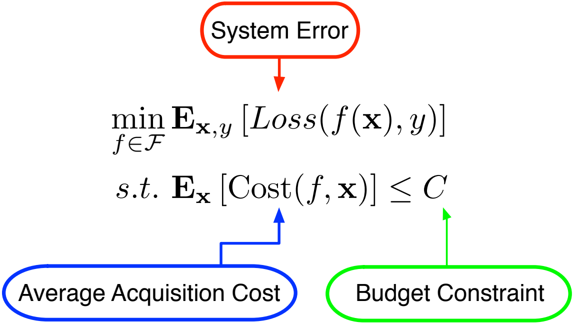
- Sensors:
  - physical measurement in some modalities
  - or computing features of various complexity
- Cost: resources, time, computation ...
- feature=measurement (possibly high dimensional)

Cost Sensitive Objective

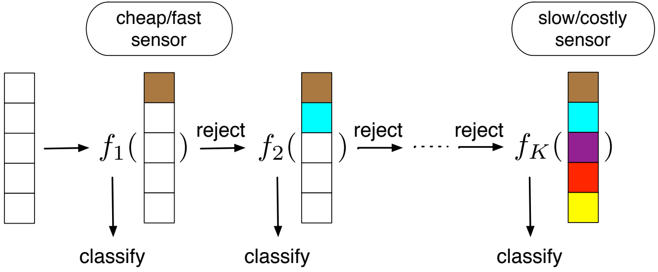
- Classifier: f
- Sample: $\mathbf{x} = [x_1 \ x_2 \ \ldots x_K]$, True label: y
- Cost of using f: $\text{Cost}(f, \mathbf{x}) = \sum_k \delta_k \mathbb{1}_{[f(\mathbf{x}) \text{ uses feature } k]}$
- Objective:

$$\min_{f \in \mathcal{F}} \mathbf{E}_{\mathbf{x},y} \left[Loss(f(\mathbf{x}), y) \right]$$

$$s.t. \ \mathbf{E}_{\mathbf{x}} \left[\text{Cost}(f, \mathbf{x}) \right] \leq C$$



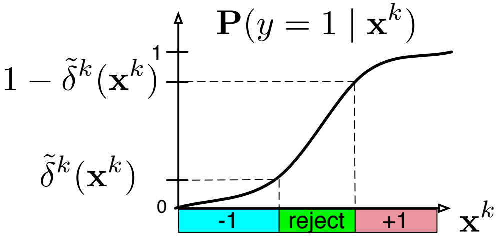
Multi-Stage Decision System (Our work)



- Assume order of stages/sensors is fixed
- Sample: $\mathbf{x} = [x_1 \ x_2 \ \ldots x_K]$, True label: y
- kth stage:
  - acquires kth feature for a cost δ_k
  - $f_k(\mathbf{x}^k)$: full decision with a reject option
  - $\mathbf{x}^k$: first k features of $\mathbf{x}$

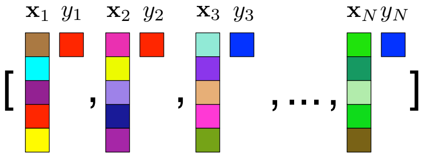
Our Approach

- **1.** Define System Risk: $= \sum_k$ Stage k Risk
  - Conditioned on: $\mathbf{x}$ is still active at kth stage
  - Stage k Risk $= \begin{cases} \delta_{k+1} & \text{, if rejects to next stage} \\ 1 & \text{, if stage } k \text{ misclassifies and not rejects} \end{cases}$
- **2.** Derive Optimal Solution if prob. distr. are given
  - Dynamic Program
  - Reduces to single stage optimization if cost-to-go is known
    - Cost-to-go, $\tilde{\delta}^k(\mathbf{x}^k) =$ expected risk of later stages $+\delta_{k+1}$



Our Approach (con'd)

- **3.** Mimic Optimal Solution in the empirical setting
  - Given training data with full features:



  - At each stage formulate:
    - Empirical risk
    - Empirical estimate of cost-to-go
  - Classifier with reject option
    - Parametrize in a convenient manner
    - Reduce to a series of supervised learning problems
  - Cyclic optimization over one stage at a time

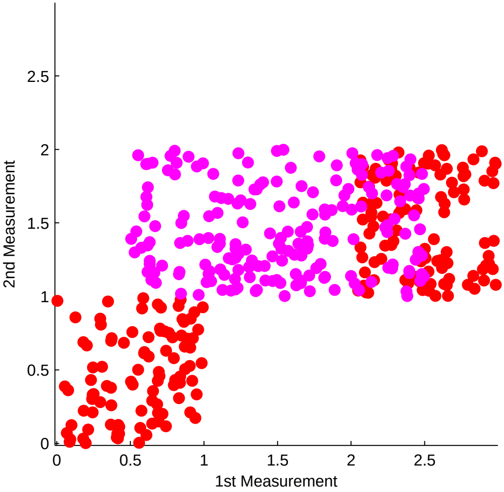
Alternative approach: single stage design of classifiers

- Myopic approach, at each stage k
  - Reject a constant fraction to next stage
  - Ignores performance of stages $k+1 \;\ldots\; K$.

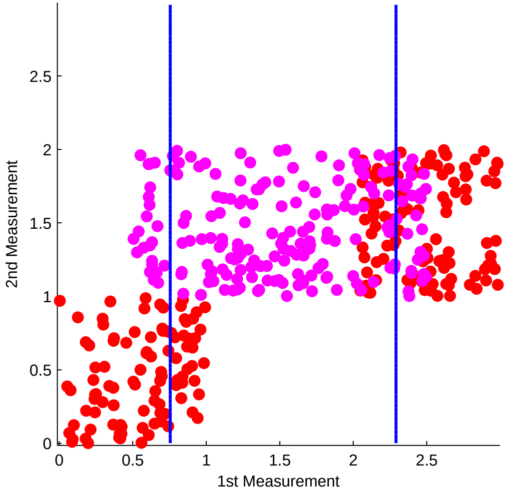
$$\text{Decision at } k\text{th stage} = \begin{cases} \text{classify}, & \text{confidence} \leq \text{threshold} \\ \text{reject to next stage}, & \text{confidence} > \text{threshold} \end{cases}$$

- Our Approach,
  - Takes the risk of the entire system into account

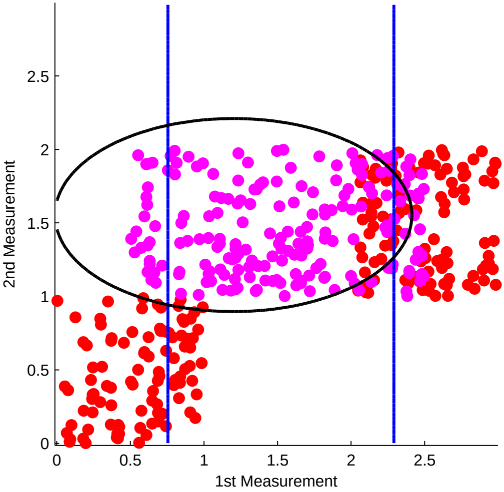
Synthetic Example



Synthetic Example: 1st Stage Classifier

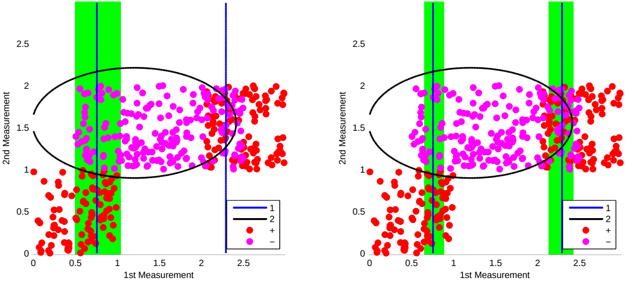


Synthetic Example: 2nd Stage Classifier



Synthetic Example: Ours vs. Myopic

Figure : Constant Budget = .3



(a) Ours: Error = .148

(b) Myopic: Error = .19

Our approach achieves smaller error for the same budget

Evaluating Performance

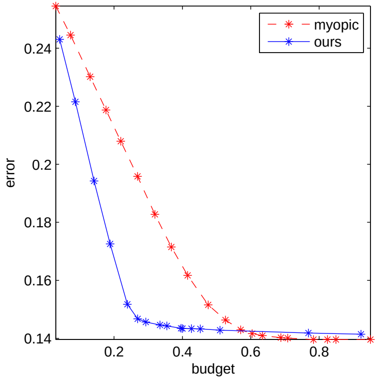
- Metrics:
  - System Test Error $=$ Error of $\mathbf{x}_i$'s classified at 1st stage $+$ Error of $\mathbf{x}_i$'s classified at 2nd stage$+$ ... $+$
  - Test Budget$=$Average Acquisition Cost per $\mathbf{x}_i$
- Operating Points
  - Ours: sweep trade-off parameter (error vs cost)
  - Myopic: sweep fraction rejected at a stage

Synthetic Example: Error vs Budget

Stage	Sensor	Cost
1	1st dim	0
2	2nd dim	1

t

For all budgets, our approach has overall better performance than myopic



MNIST (UCI)

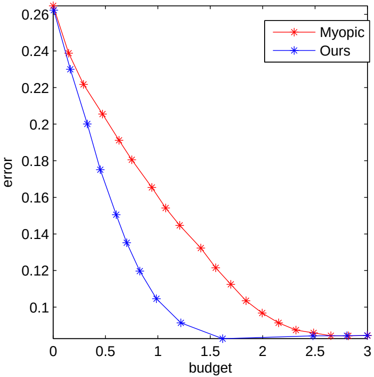
- $\mathbf{x}$ = Handwritten image of a digits
- y: 1 of 10 digits

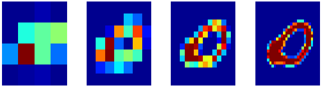
t

Stage	Sensor Resolution	Cost
1	4x4	0
2	7x7	1
3	14x14	2
4	28x28	3

- Full resolution: cost=3

Can achieve full resolution performance with low resolution measurements



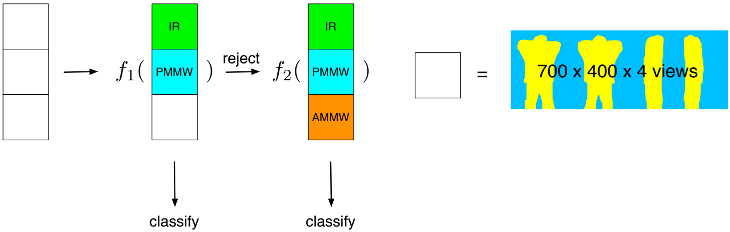


Concealed Explosive Detection Data

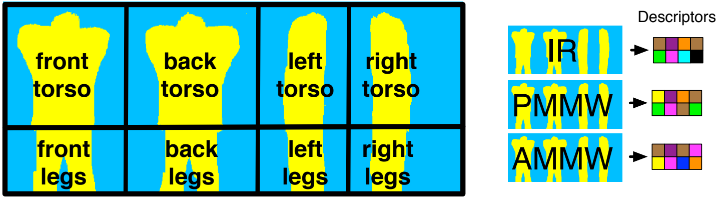
- Standoff images of subjects (people) wearing explosive devices underneath clothing
- Dataset Statistics

# of Samples	1230
Modalities	IR, PMMW, AMMW
# of Views	4
Image Size/View	700x400

- Several types of threats (vest bombs, etc)
- 70% threats, 30% clean
- Classification objective: is subject concealing a threat?



Our Method



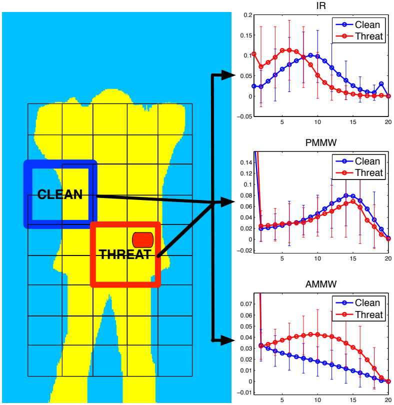
1. Divide Body into 8 regions
2. Reduce dimensionality per modality
    - Find a confidence for each region
    - 700x400x4 $\rightarrow$ 8 dimensional descriptor x 3 modalities
3. Use low dim. descriptor as input to our system

Test our approach using simple pre-processing

Extract Overlapping Windows

- For a window
  - 20 bins of normalized pixel intensity
  - compute histogram of pixel values

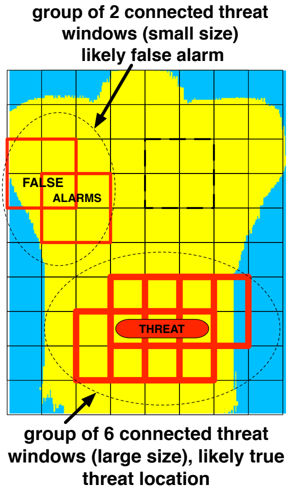
- AMMW: best differentiator
- IR and PMMW: worse



Descriptor for Each Region

t

1. Learn a window classifier
   - threat or clean
   - for each modality: IR, PMMW, AMMW
2. Evaluate each window in a region
3. Find connected threat windows
4. Report the size of the largest group
   - Descriptors: 700x400x4 → 8
   - Input to our system

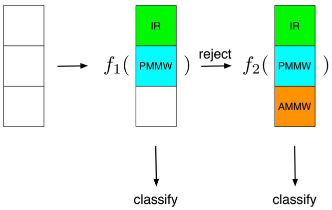


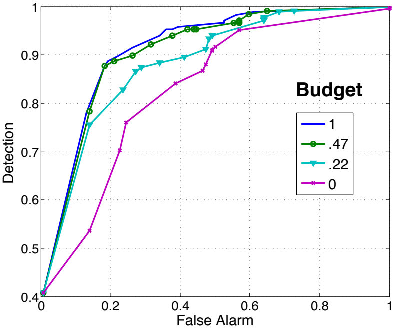
ROC for varying budget

- Split dataset: 50% train, 50% test
- **x** = confidence vector per sensor
- $y \in \{\text{Threat}, \text{Not Threat}\}$
- Better pre-processing will improve baseline performance

t

Stage	Sensor	Cost
1	IR,PMMW	0
2	AMMW	1





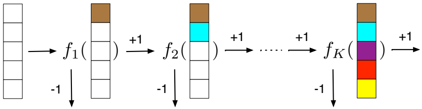
Can achieve near-optimal performance using expensive sensor less than half the time!

Conclusion

- Developed a theory for designing non-parametric multi-stage multi-class classifiers
- Can be adapted to extend existing machine learning approaches
- Future Work:
  - Optimize sequencing of sensors when choice is possible
  - Explore alternatives
- This work appears in:
  - K. Trapeznikov, V. Saligrama, D. Castañón, *Multi-stage Classifier Design*, Asian Conference on Machine Learning, 2012
  - K. Trapeznikov, V. Saligrama, D. Castañón, *Two Stage Decision System*, IEEE Statistical Signal Processing, 2012

Related Work

- Parametric Methods (estimate/model $\mathbf{P}(\mathbf{x}, y)$ or transition probabilities $\mathbf{P}(x_1 \mid x_2)$)
  - Markov Decision Process: [Ji and Carin, 2007, Kapoor and Horvitz, 2009]
  - Decision Tree based: [Sheng and Ling, 2006, Bilgic and Getoor, 2007, Zubek and Dietterich, 2002]
  - Entropy Maximizing: [Kanani and Melville, 2008].
- Non-parametric methods
  - Detection Cascades ([Viola and Jones, 2001, Chen et al., 2012])
    - Partially-Adaptive, reduce acquisition cost for one class
    - Partial Decisions at each stage
    - No multi-class extensions



  - Myopic Aproaches ([Liu et al., 2008])
    - Ignorant of performance later stages

References

Bilgic, M. and Getoor, L. (2007).
Voila: Efficient feature-value acquisition for classification.
In *AAAI*.

Chen, M., Xu, Z., Weinberger, K. Q., Chapelle, O., and Kedem, D. (2012).
Classifier cascade: Tradeoff between accuracy and feature evaluation cost.
In *AISTATS*.

Ji, S. and Carin, L. (2007).
Cost-sensitive feature acquisition and classification.
In *Pattern Recognition*.

Kanani, P. and Melville, P. (2008).
Prediction-time active feature-value acquisition for cost-effective customer targeting.
In *NIPS*.

Kapoor, A. and Horvitz, E. (2009).
Breaking boundaries: Active information acquisition across learning and diagnosis.
In *NIPS*.

Liu, L.-P., Yu, Y., Jiang, Y., and Zhou, Z.-H. (2008).
Tefe: A time-efficient approach to feature extraction.
In *ICDM*.

Sheng, V. S. and Ling, C. X. (2006).
Feature value acquisition in testing: A sequential batch test algorithm.
In *ICML*, pages 809–816.

Viola, P. and Jones, M. (2001).
Robust real-time object detection.
In *Int. J. of Comp. Vis.*

Zubek, V. B. and Dietterich, T. G. (2002).
Pruning improves heuristic search for cost-sensitive learning.