Towards Efficient Learning of Optimal Spatial Bag-of-Words Representations

Lu Jiang[1], Wei Tong[1], Deyu Meng[2], Alexander G. Hauptmann[1]

[1] School of Computer Science, Carnegie Mellon University

[2]School of Mathematics and Statistics, Xi'an Jiaotong University

Carnegie Mellon University

People

CMU Informedia Team

Wei Tong

Deyu Meng

Alexander G. Hauptmann

Carnegie Mellon University

Outline

- Motivation
- Related Work
- Jensen - Shannon Tiling
- Experiment Results
- Conclusions

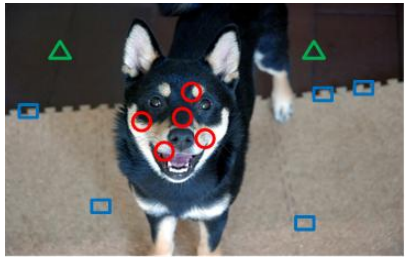
Carnegie Mellon University

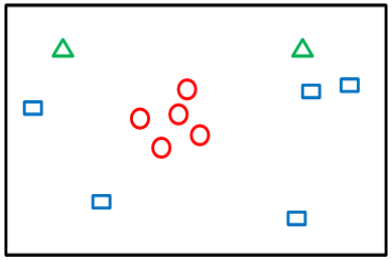
Outline

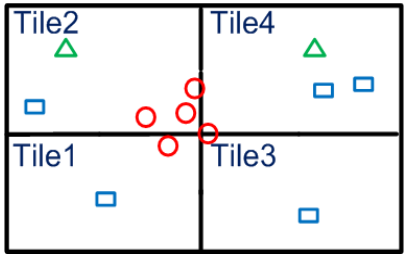
- Motivation
- Related Work
- Jensen - Shannon Tiling
- Experiment Results
- Conclusions

Spatial Bag-of-Words

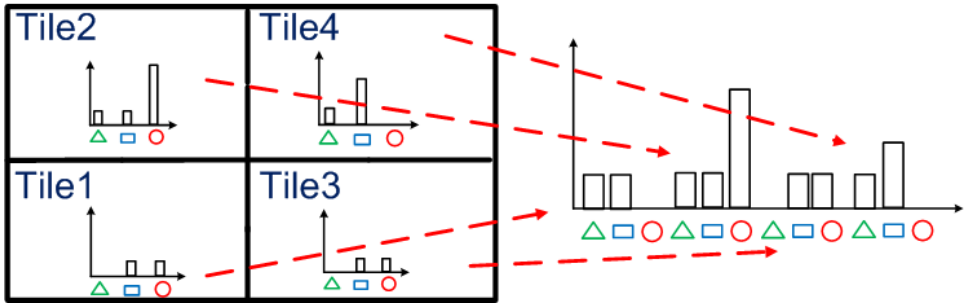
- The Spatial Bag-of-Words (BoW) model has proven one of the most broadly used models in image and video retrieval.
- It divides an image/video into one or more smaller tiles.







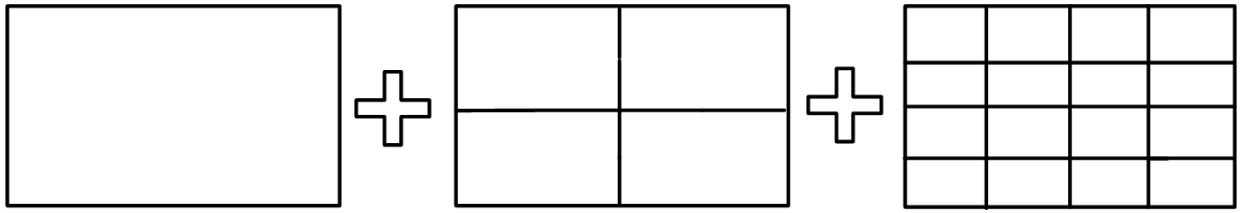
- The image represented by the concatenated BoW histograms from all the tiles.



Carnegie Mellon University

Spatial Pyramid Matching (SPM)

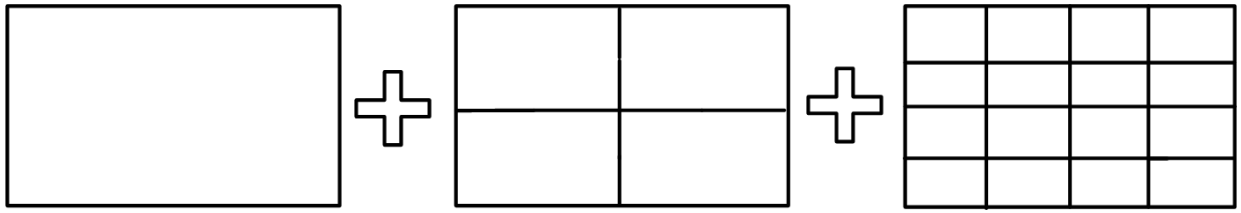
- Spatial Pyramid Matching is a robust extension to spatial BoW Model.
- Combine a set of predefined partitions (1x1, 2x2, 4x4, etc.)



- But, are predefined representations in SPM sufficient for multimedia retrieval?

Spatial Pyramid Matching (SPM)

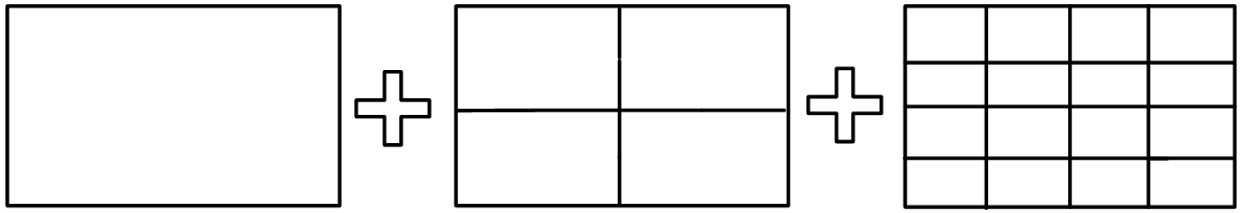
- Spatial Pyramid Matching is a robust extension to spatial BoW Model.
- Combine a set of predefined partitions (1x1, 2x2, 4x4, etc.)



- But, are predefined tilings in SPM sufficient for multimedia retrieval?

Spatial Pyramid Matching (SPM)

- Spatial Pyramid Matching is a robust extension to spatial BoW Model.
- Combine a set of predefined partitions (1x1, 2x2, 4x4, etc.)

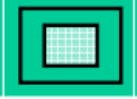
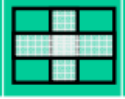
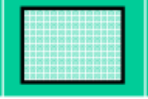
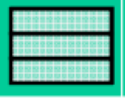
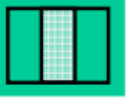


- But, are predefined representations SPM sufficient for multimedia retrieval?

IBM's Talk @ TRECVID 12

Semantic Indexing

Global Visual Features - Spatial Granularities

	Center	Cross	Global	Grid	Horizontal	Horiz. Parts	Layout	Vertical
								
Color Correlogram	X	X	X		x		X	X
Color Histogram	X	X	X		X	X	X	X
Color Moments	X		X			x		X
Color Wavelet		x	X					
Color Wavelet Texture	X		X		X	x	X	X
Fourier Polar Pyramid	X		X					
Edge Histogram	X		X		X	X	X	x
GIST			X					
Image Stats			X	X				
Image Type	X		x	X	X	x		x
LBP histogram			X					
Maxi Thumbnail Vector			X					
Mini Thumbnail Vector	X		X					
Siftogram			X					
Size Vector			X					
Thumbnail Vector	X		X					
Wavelet Texture	X		X					
Curvelet Texture			x	x				

http://www-nlpir.nist.gov/projects/tvpubs/tv12.slides/tv12.ibm.sin.slides.pdf

SRI Sarnoff's Talk @TRECVID 12

Multimedia Event Detection

Feature Pooling Using Fixed Spatial Patterns

- Objective
  - Limitation: Features aggregated from a whole frame contains more irrelevant data of an event
  - Goal: Extract event relevant information by pooling features from different parts of a frame
- Spatial pooling using fixed patterns
  - Aggregate features over a set of pre-defined regions as shown at
  - Implicitly encodes location information with visual-words for bet
  - Fixed patterns are easy and fast to computer

SRI International SARNOFF

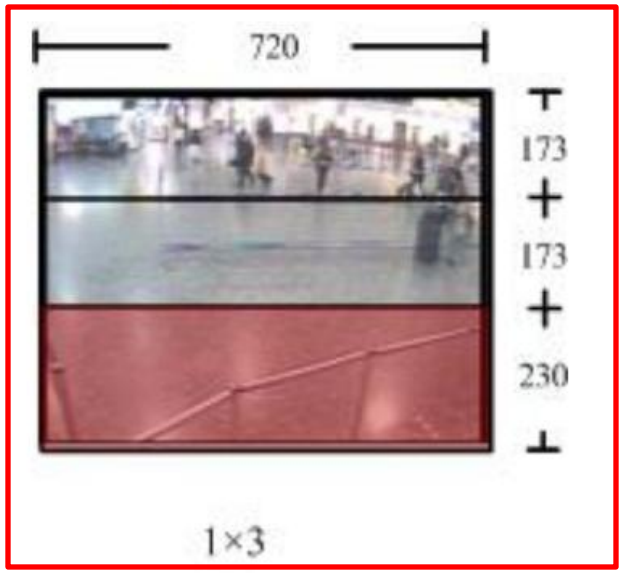
5

http://www-nlpir.nist.gov/projects/tvpubs/tv12.slides/tv12.aurora.med.slides.pdf

CMU's Talk @ TRECVID 11

Surveillance Event Detection

- Each frame is divided into a set of rectangular tiles or grids.
- The resulting Bow features are derived by concatenating the BoW features captured in each grid.
- Encode the adjusted spatial information in BoW.



http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.cmu.sed.slides.pdf

Carnegie Mellon University

Motivation

- Spatial Representation is **fundamental** to multimedia retrieval.
  - Semantic objects/concepts indexing.
  - Multimedia event retrieval.
  - Surveillance event detection, etc.
- Different spatial representations can affects results considerably.

Carnegie Mellon University

Semi-Manual Approach

- A straightforward way to find optimal representations [1,2]:
  - Manually design representation candidates.
  - Verify the candidates by running the classifier.
- Cons:
  - Require manual effort .
  - **Computationally infeasible** to verify all the candidates.

[1] W. Tong, Y. Yang, L. Jiang, S. I. Yu, Z. Lan, Z. Ma, W. Sze, E. Younessian, and A. G. Hauptmann. E-LAMP: integration of innovative ideas for multimedia event detection. Machine Vision and Applications, pages 1–11, 2013.
[2] V. Viitaniemi and J. Laaksonen. Spatial extensions to bag of visual words. In ACM CIVR, 2009.

Carnegie Mellon University

Motivation

- Manually designing representations is never an easy thing.
- Our goal:
  - Automatically learn salient spatial representations from data.
  - Efficient enough to run on large-scale data.

Carnegie Mellon University

Outline

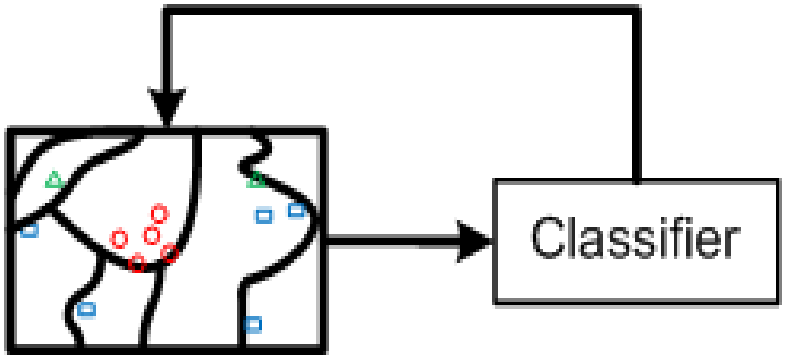
- Motivation
- Related Work
- Jensen - Shannon Tiling
- Experiment Results
- Conclusions

Comparison with Related Work

Existing studies learn the representations with the classifiers [3,4,5] .

- Reasonable Improvements.
- Time consuming.
- Low cost-effective.
- 2,000 core hours for 2% MAP

(**worth doing?**)



[3] J. Feng, B. Ni, Q. Tian, and S. Yan. Geometric lp-norm feature pooling for image classification. In CVPR, 2011.
[4] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In CVPR, 2012.
[5] G. Sharma and F. Jurie. Learning discriminative spatial representation for image classification. In BMVC, 2011.

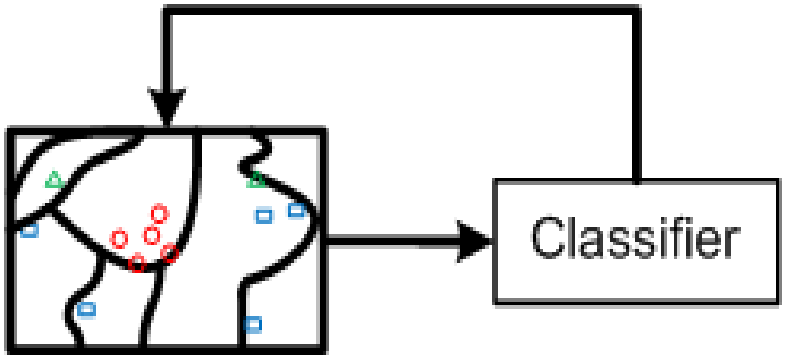
Comparison with Related Work

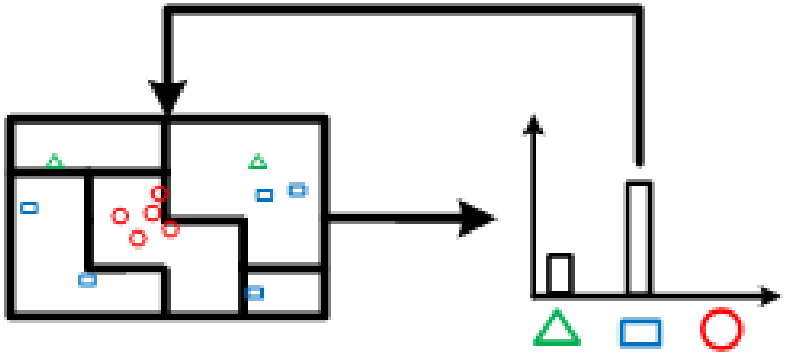
Existing studies learn the representations with the classifiers [3,4,5] .

- Reasonable Improvements.
- Time consuming.
- Low cost-effective.
- 2,000 core hours for 2% MAP (worth doing?)

JS(Jensen-Shannon)- Tiling directly captures representations at lower BoW level, **independent of the classifier**.

- Decent improvements.
- Orders of magnitude faster.
- **High cost-effective.**





BoW Distribution

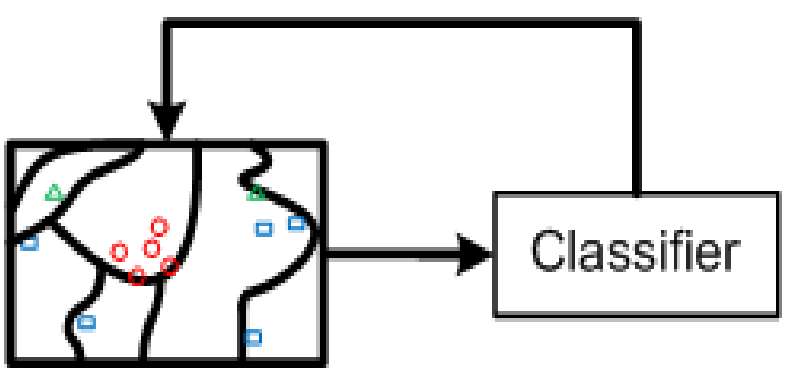
Comparison with Related Work

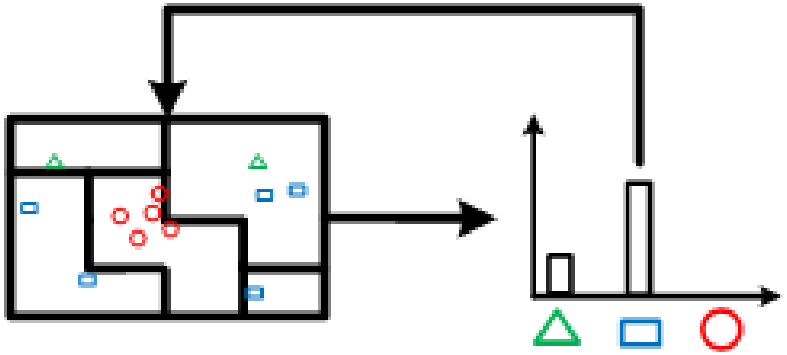
Existing Work learn the representations with the classifiers [3,4,5] .

- **Embedded** method in feature selection.

JS Tiling directly captures them at lower BoW level, independent of the classifier.

- **Filter** method in feature selection.
- **Efficiency.**
- **Generalizability.**





BoW Distribution

Carnegie Mellon University

Proposed Approach

- **JS(Jensen-Shannon)-Tiling** offers a solution because it is:
  - Learn salient representations automatically from data.
  - Applicably to large-scale datsets.
- It is **an important component** in CMU Teams' final submission in TRECVID 2012 Multimedia Event Detection[1].

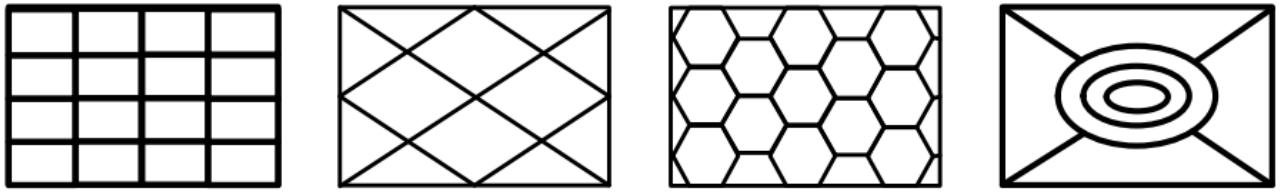
Carnegie Mellon University

Outline

- Motivation
- Related Work
- Jensen - Shannon Tiling
- Experiment Results
- Conclusions

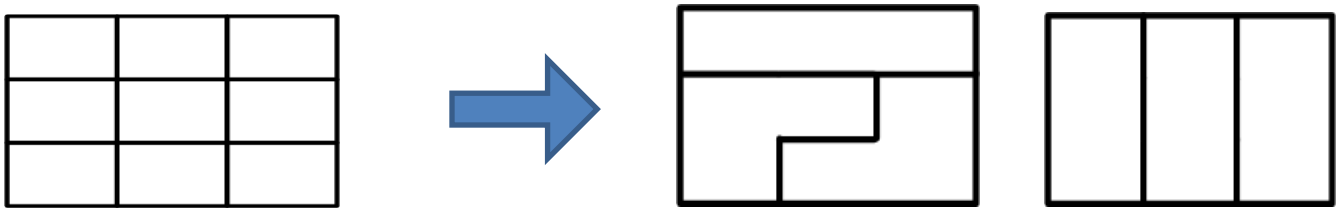
Problem Formulation

- A mask is a predefined partition.



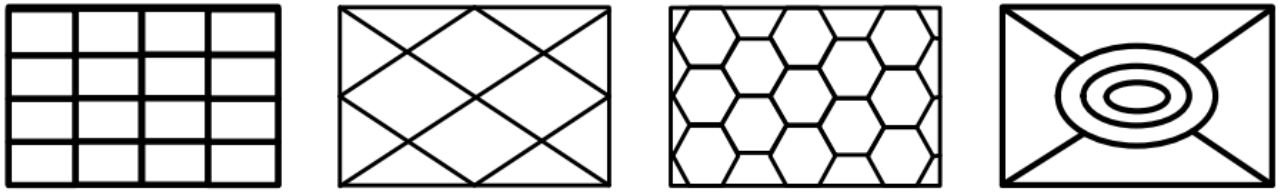
(a) rectangle (b) diamond (c) hexagon (d) ellipse

- More representations can be derived by combining the tiles in the mask.
- Each representation is called a tiling.



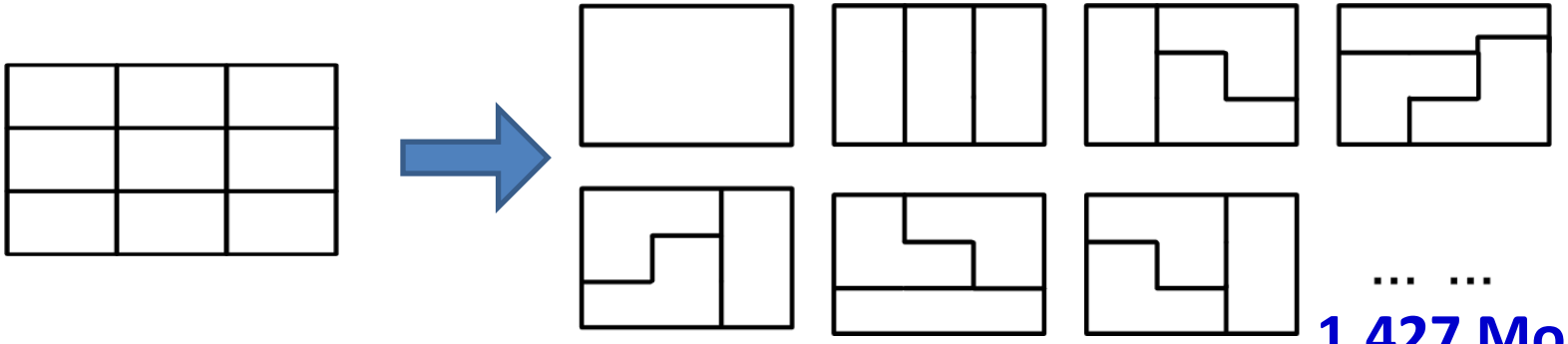
Problem Formulation

- A mask is a predefined partition.



(a) rectangle (b) diamond (c) hexagon (d) ellipse

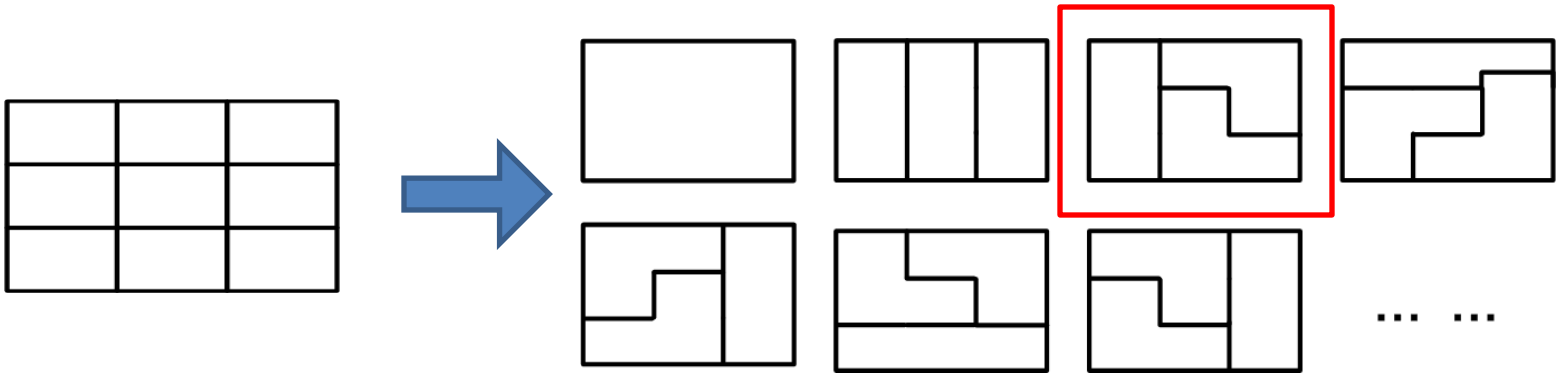
- More representations can be derived by combining the tiles in the mask.
- Each representation is called a tiling.



… …

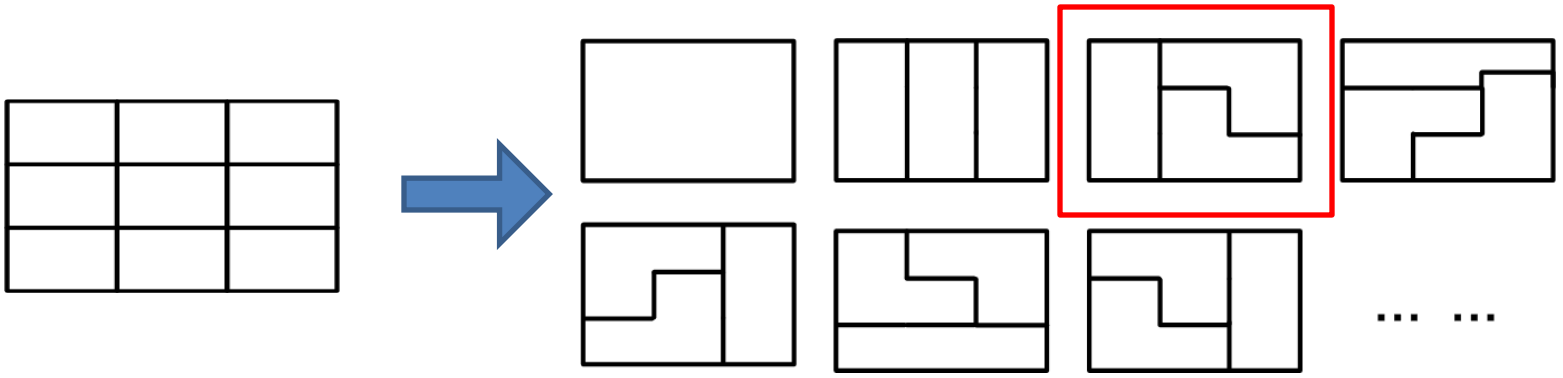
1,427 More

Problem Formulation



- Problem: Find optimal tilings for a given mask.
- Proposed approach:
  - Systematically generate all possible tilings from the given mask.
  - Efficiently evaluate each tiling without running classifiers.

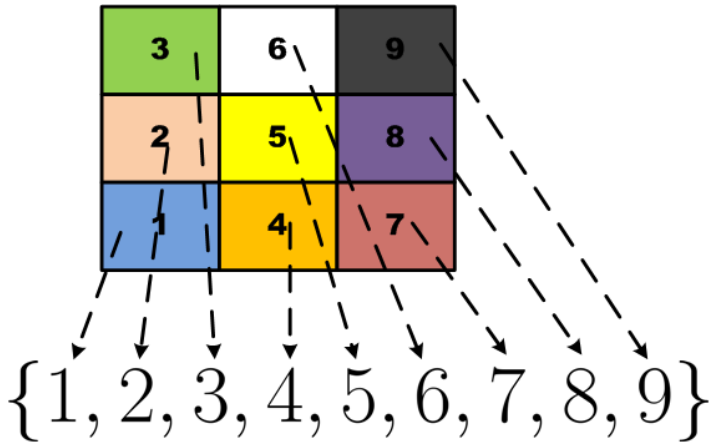
Problem Formulation



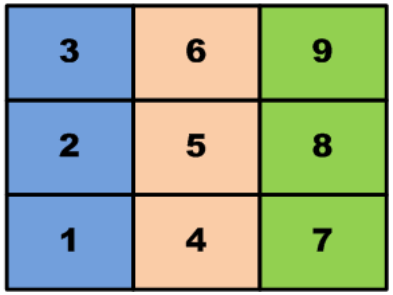
- Problem: Find optimal tilings for a given mask.
- Proposed approach:
  - **Systematically generate all possible tilings from the given mask.**
  - Efficiently evaluate each tiling without running classifiers.

Tiling Definition

- Tiling can be defined based on the set-partition theory.
- Divide a set as a union of non-overlapping and non-empty subsets.

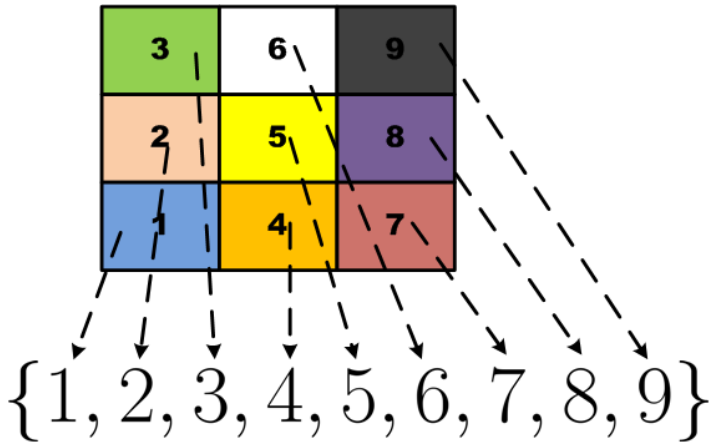


$$\{\{1,2,3\},\{4,5,6\},\{7,8,9\}\}$$

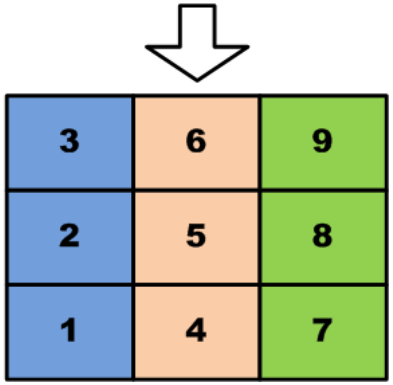


Tiling Definition

- Tiling can be defined based on the set-partition theory.
- Divide a set as a union of non-overlapping and non-empty subsets.



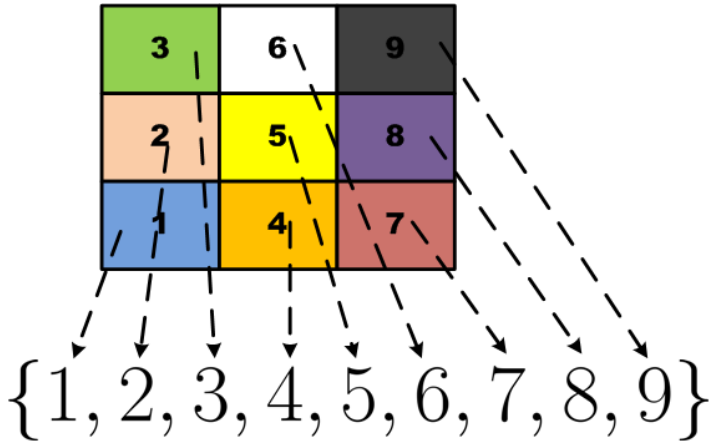
$$\{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}\}$$



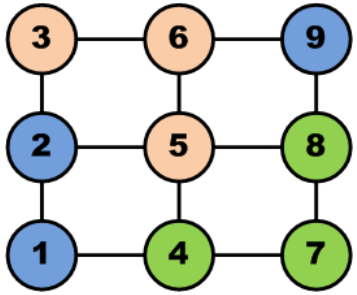
- A tiling can be defined as:
  - A complete partition of mask into non-overlapping area.
  - Each partition (tile) is visually adjacent[3].

Tiling Definition

- Tiling can be defined based on the set-partition theory.
- Divide a set as a union of non-overlapping and non-empty subsets.



(c) Not a tiling.



identical to the connected components in the graph.

- A tiling can be defined as:
  - A complete partition of mask into non-overlapping area.
  - Each partition (tile) is visually adjacent[3].

Tiling Generation

NP-hard problem. But given reasonable masks, it is solvable.

Algorithm (Loop until termination):

1) Generate a set partition candidate;
2) Test whether this candidate obeys the adjacency constraint;

Type	Parameter	#Set Partition	#Tiling	#Equal Tiling
Rectangle	2×2	15	12	4
Rectangle	3×3	21147	1434	12
Rectangle	4×4	10480142147	1691690	225
Diamond	1×1	15	12	4
Diamond	2×2	52	16	2
Diamond	3×3	4213597	17326	23
Hexagon	1	52	20	2
Hexagon	1.5	4140	466	7
Ellipse	4	4140	344	5
Ellipse	8	4213597	5504	10

- Visual adjacency constraint **significantly reduces** the number of candidates.

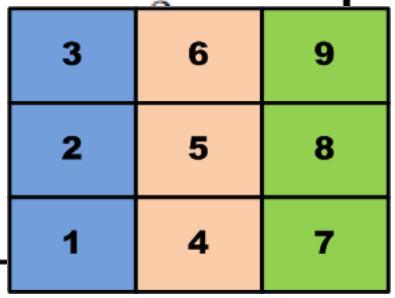
Tiling Generation

NP-hard problem. But given reasonable masks, it is solvable.

Algorithm (Loop until termination):

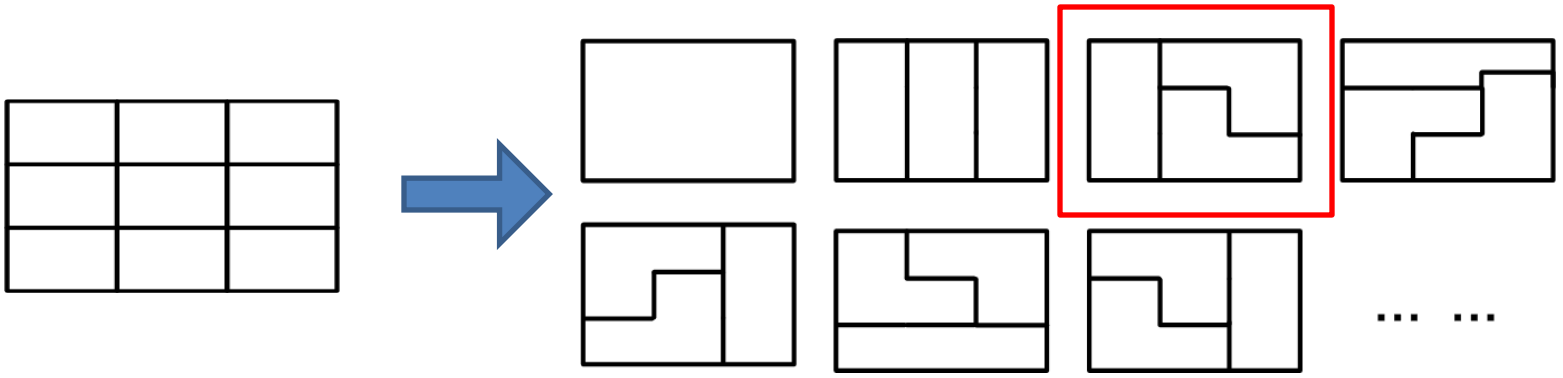
1) Generate a set partition candidate;
2) Test whether this candidate obeys the adjacency constraint;

Type	Parameter	#Set Partition	#Tiling	#Equal Tiling
Rectangle	2×2	15	12	4
Rectangle	3×3	21147	1434	12
Rectangle	4×4	10480142147	1691690	225
Diamond	1×1	15	12	4
Diamond	2×2	52	16	
Diamond	3×3	4213597	17326	
Hexagon	1	52	20	
Hexagon	1.5	4140	466	
Ellipse	4	4140	344	
Ellipse	8	4213597	5504	



- Visual adjacency constraint **significantly reduces** the number of candidates.

Problem Formulation



- Problem: Find optimal tilings for a given mask.
- Proposed approach:
  - Systematically generate all possible tilings from the given mask.
  - **Efficiently evaluate each tiling without running classifiers.**

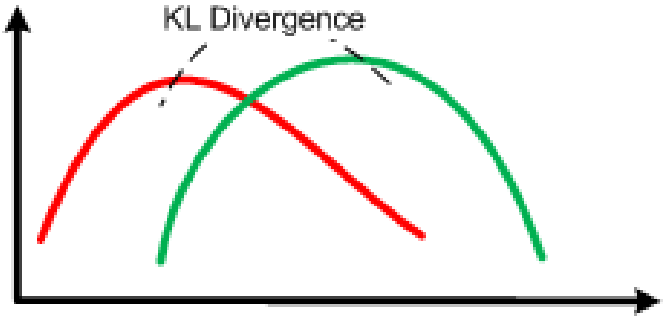
Tiling Evaluation

- Intuitively an optimal tiling would separate the positive and negative samples with the maximum distance.
- The distance is evaluated w.r.t Kullback-Leibler (KL) divergence.
- Symmetric version called Jensen-Shannon (JS) divergence.

$$cost(\mathcal{T}_\kappa) = \lambda|\mathcal{T}_\kappa(S)| - \sum_{i=0}^{|\mathcal{T}_\kappa(S)|-1} \frac{JS(D_i^+ \parallel D_i^-)}{|\mathcal{T}_\kappa(S)|}$$

$\mathcal{T}_\kappa(S)$ is the tiling to evaluate.

D_i^+ D_i^- average word distributions of positive and negative samples generated by the tiling.



Carnegie Mellon University

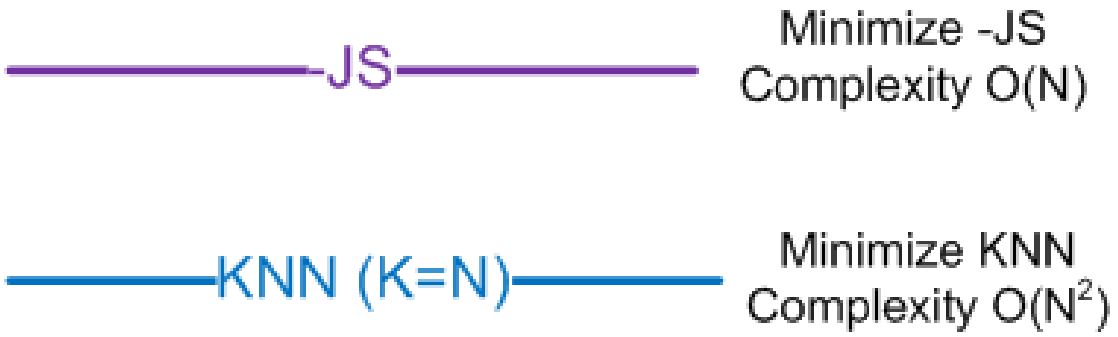
Tiling Evaluation

- Consistent with the distribution separability principle in [6].

[6]Y. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In ICML, 2010.

Tiling Evaluation

- Consistent with the distribution separability principle in [6].
- We prove that the negative JS divergence is approximately **an upper bound** of the training error of a weighted K-Nearest Neighbor classifier K = N.
- Justify why the computationally inexpensive divergence can be a proxy to the computationally expensive classifier.



Carnegie Mellon University

Outline

- Motivation
- Related Work
- Jensen - Shannon Tiling
- Experiment Results
- Conclusions

Carnegie Mellon University

Comparison with state-of-the-art

Dataset	Method	MAP	Accuracy
15-Scene	SPM [12]	83.5±0.5	80.8±0.6
	Boureau et al. [2]	-	84.9±0.3
	Sharma et al. [19]	85.5±0.7	-
	van Gemert et al. [23]	-	76.7±0.4
	Sharma et al. [18]	-	81.2±0.6
	Yang et al. [27]	-	80.3±0.9
	JS Tiling	**88.0±0.3**	**85.3±0.4**
	Method	MAP	Min DCR
SED	SPM [12]	22.8±1.0	89.0±1.5
	Winner'11 [30]	23.8±0.8	87.2±1.0
	JS Tiling	**26.5±0.6**	**85.1±0.9**
	Method	MAP(SIFT)	MAP(STIP)
MED	SPM [12]	26.8	17.2
	Winner'12 [29, 21]	27.3	18.7
	JS Tiling	**30.7**	**21.2**
	Method	MAP	-
VOC	SPM [12]	52.5	-
	Winner'07 [15]	54.2	-
	Wang et al. [26]	55.1	-
	Yang et al. [28]	**59.6**	-
	JS Tiling	55.5	-

- **Consistently outperforms the SPM** across datasets on scene/object recognition and event detection.
- **Comparable or even better** results with existing methods.

Reasons for the Improvement

- 1) Capture more **salient spatial representations** than SPM.

Rank	Predefined Masks			Rectangle Masks			All Masks		
	Tiling	Accuracy	MAP	Tiling	Accuracy	MAP	Tiling	Accuracy	MAP
1		79.5±0.7	81.5±0.6		80.4±0.7	83.2±0.6		82.4±0.4	85.5±0.4
2		79.4±0.6	81.8±0.6		80.4±0.4	83.0±0.6		81.4±0.4	84.3±0.5
3		78.6±0.4	80.7±0.4		80.0±0.6	82.4±0.5		80.8±0.5	83.7±0.5
4		77.5±0.2	80.3±0.4		79.9±0.5	82.1±0.7		80.9±0.3	82.5±0.4
5		77.8±0.5	79.6±0.5		79.5±0.7	81.5±0.6		80.4±0.7	83.2±0.6

Predefined tilings in SPM

Proposed Method

The results are on 15 scene category dataset.

Carnegie Mellon University

Reasons for the Improvement

- 1) Capture more salient spatial representations than SPM.

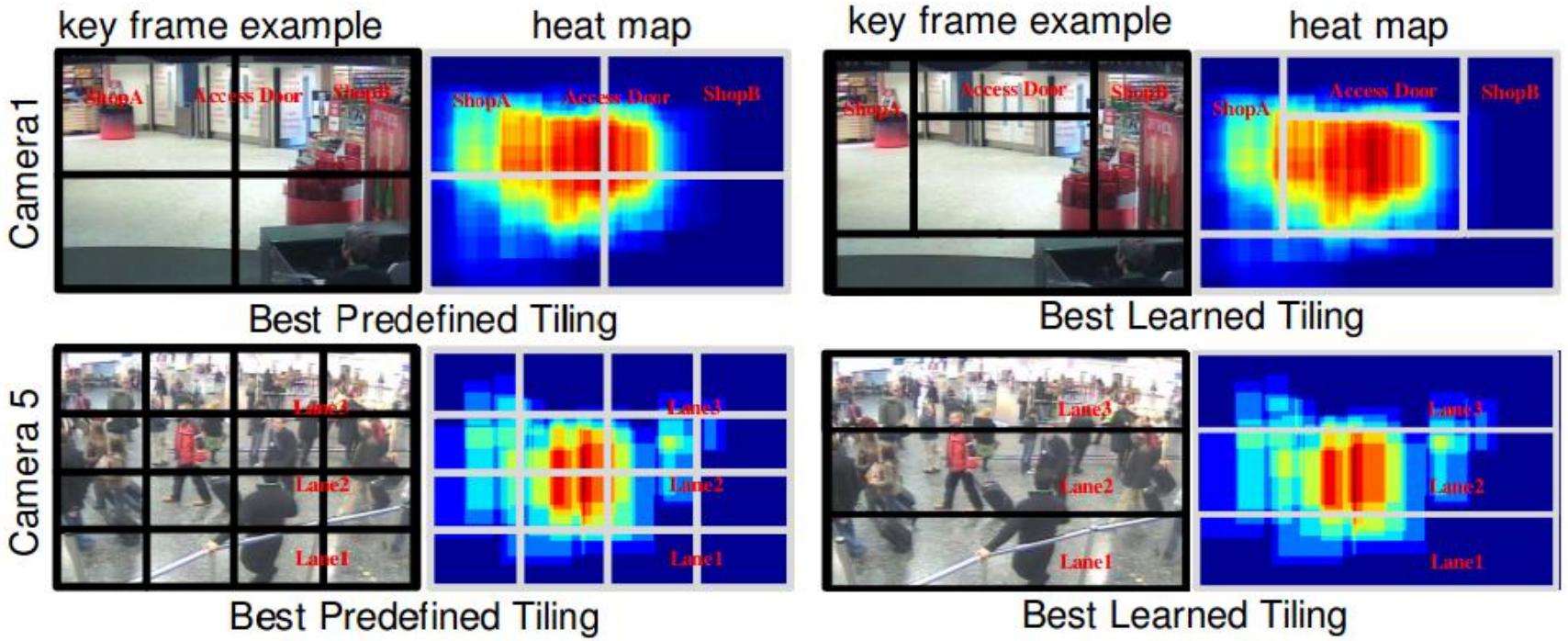
Rank	Predefined Masks			Rectangle Masks			All Masks		
	Tiling	Accuracy	MAP	Tiling	Accuracy	MAP	Tiling	Accuracy	MAP
1		79.5±0.7	81.5±0.6		80.4±0.7	83.2±0.6		82.4±0.4	85.5±0.4
2		79.4±0.6	81.8±0.6		80.4±0.4	83.0±0.6		81.4±0.4	84.3±0.5
3		78.6±0.4	80.7±0.4		80.0±0.6	82.4±0.5		80.8±0.5	83.7±0.5
4		77.5±0.2	80.3±0.4		79.9±0.5	82.1±0.7		80.9±0.3	82.5±0.4
5		77.8±0.5	79.6±0.5		79.5±0.7	81.5±0.6		80.4±0.7	83.2±0.6

- 2) Substantially **augment the choices of representations**.

L	Spatial Pyramid		Rectangle Masks		All Masks	
	Accuracy	MAP	Accuracy	MAP	Accuracy	MAP
0	75.3±0.3	81.5±0.6	80.4±0.7	83.2±0.6	82.4±0.4	85.5±0.4
1	80.7±0.6	83.3±0.6	80.8±0.5	83.6±0.6	82.2±0.5	85.4±0.4
2	**80.8±0.6**	**83.5±0.5**	81.4±0.6	84.1±0.6	82.7±0.6	85.8±0.4
3	80.1±0.6	82.4±0.5	81.5±0.6	84.1±0.7	82.8±0.5	85.8±0.4
4	79.2±0.6	81.2±0.6	81.7±0.6	84.2±0.6	83.5±0.7	86.7±0.5
7	-	-	**81.9±0.5**	**84.6±0.5**	**85.3±0.4**	**88.0±0.3**

The results are on 15 scene category dataset.

Learned Tiling on SED dataset



- Heat maps are plotted based on manual annotations.
- Tilings are learned without using annotations.
- Learned tilings are **more sensible** than predefined tilings.

Carnegie Mellon University

Runtime Comparison

- Compare the runtime with tiling selection by running classifiers.
- Search a space of 1,434 tilings.

Dataset	JS Tiling	Linear SVM	Kernel SVM
15-scene	1.1(h)	1,314(h)	10,874(h)
SED	2.1(h)	2,629(h)	32,862(h)
MED	2.3(h)	4,541(h)	41,825(h)
Pascal VOC	1.6(h)	1,912(h)	22,346(h)

- A single core Intel Core i7 CPU@2.8GHz with 4G memory.
- **Orders of magnitude faster** than running classifiers.
- Substantiate the theoretical complexity analysis.

Carnegie Mellon University

Outline

- Motivation
- Related Work
- Jensen - Shannon Tiling
- Experiment Results
- Conclusions

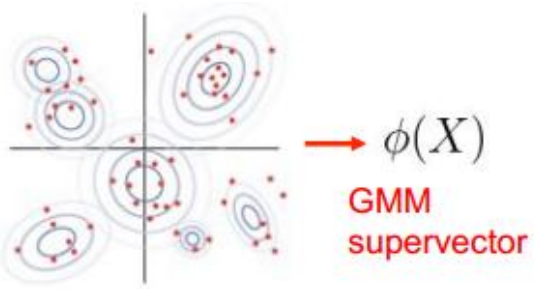
Carnegie Mellon University

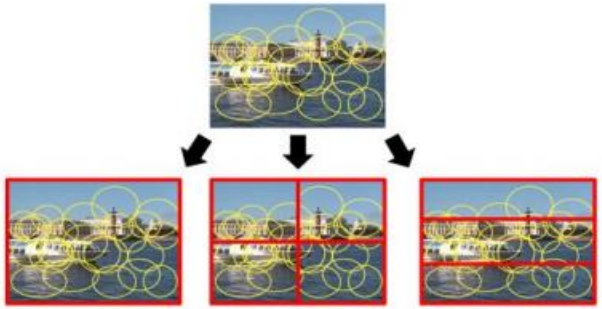
Summary

- A few messages to **take away from this talk**:
  - JS Tiling provides a efficient solution to automatically learn salient BoW representations for **large-scale datasets**.
  - JS Tiling consistently outperforms the spatial pyramid matching across datasets. Comparable or even better performance with existing methods.

Beyond BoW representation

- Tokyo TechCanon's Talk @TRECVID 2012

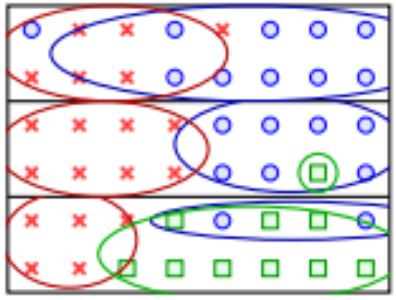




- For SIFT-Har, SIFT-Hes, HOG, SURF, and RGB-SIFT

http://www-nlpir.nist.gov/projects/tvpubs/tv12.slides/tv12.tokyotechcanon.med.slides.pdf

- AXES's Talk @TRECVID 2013
  - Spatial Fisher vector (SFV)
    (Krapac et al., ICCV, 2011)
    - encodes first and second moments of visual word locations
    - adds 6 entries for each visual word: μ and σ for (x, y, t) coordinates.
  - Compared to spatial pyramids:
    (Oneaţă et al., ICCV, 2013)
    - similar performance gain
    - SFV are more compact

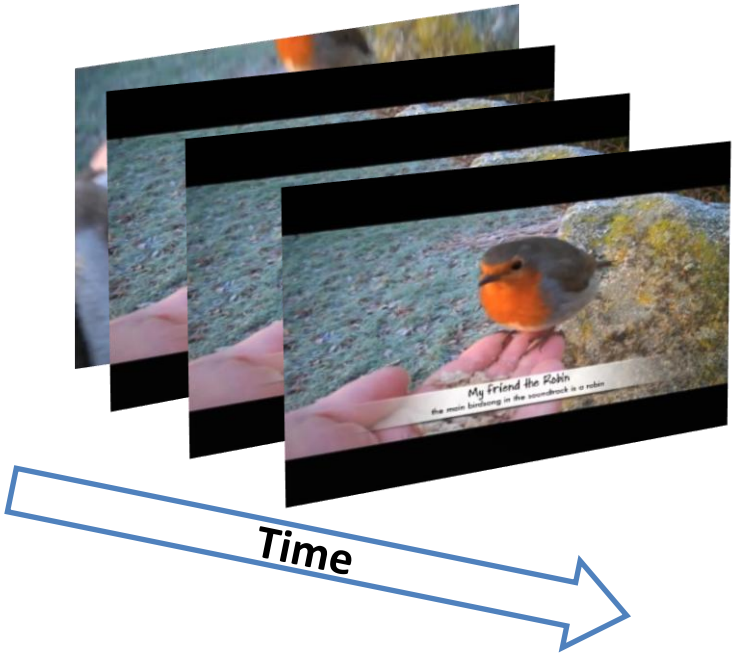


Schematic illustration of the spatial Fisher vector for three types of visual words (○, ×, □) in an image.

http://www-nlpir.nist.gov/projects/tvpubs/tv13.slides/axes.tv13.med.slides.pdf

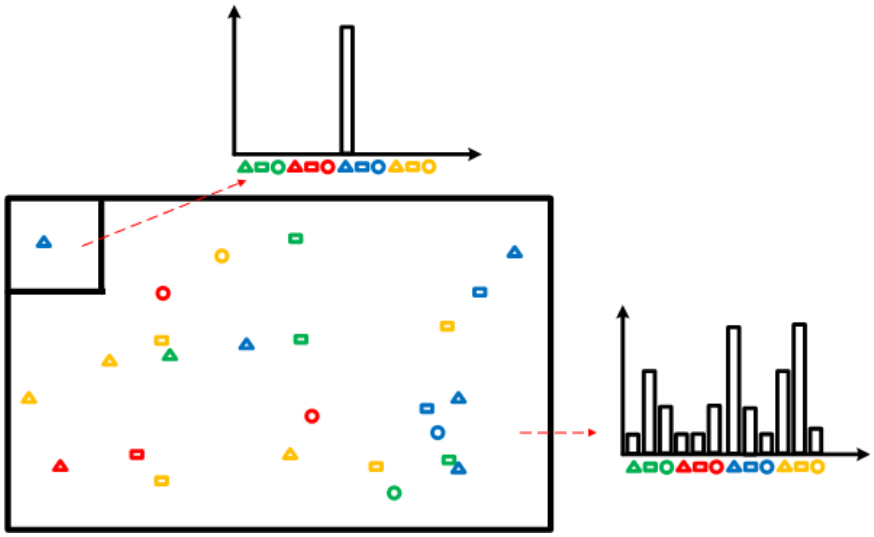
Beyond spatial representation

- Temporal tiling
  - Determine optimal sliding window sizes.



Aspects to be Improved

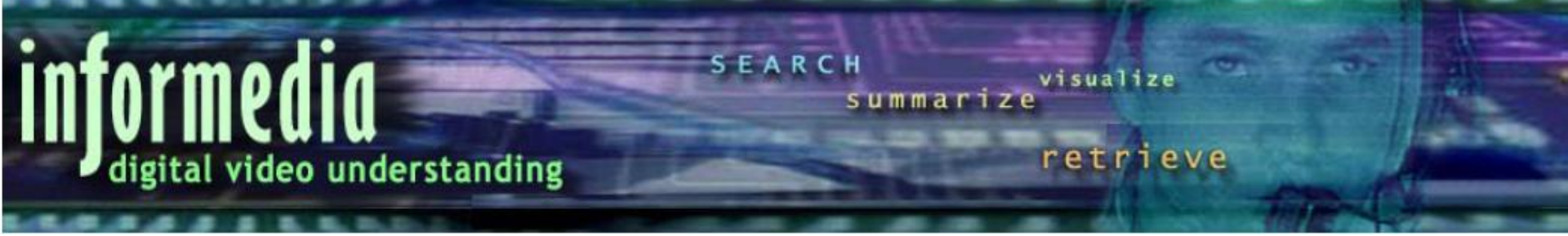
- The tilings learned from different masks are not directly comparable. A practical trick:
  - Start with a number of masks.
  - Use JS-Tiling to find a couple of salient tilings from the huge search space.
  - Run classifiers on these tilings on the validation dataset, and fuse promising ones to obtain better performance.
- Sampling bias for small tiles (overestimate the distance).
  - Equal tiling can avoid this bias.
  - Study the smoothing function.



Carnegie Mellon University

Acknowledgement

This work was partially supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.


informedia
digital video understanding
SEARCH
summarize
visualize
retrieve

THANK YOU.
Q&A?