CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

模式识别国家重点实验室
National Lab of Pattern Recognition

中国科学院自动化研究所
Institute of Automation Chinese Academy of Sciences

ACPR2015

Large-Scale Visual Computing

Tieniu Tan

Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

4 November 2015, Kuala Lumpur, Malaysia

Outline

- Background and Context
- Feature Representation
- Model Learning
- LSVC: An Example
  - Large-Scale Visual Surveillance
- Conclusions

Outline

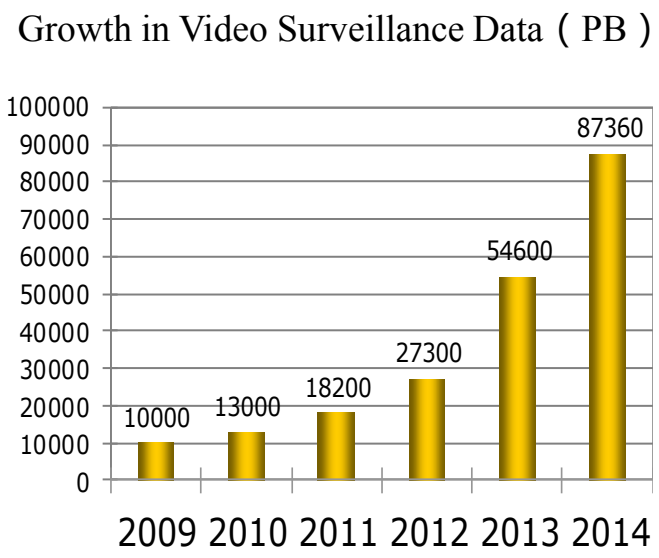
- **Background and Context**
- Feature Representation
- Model Learning
- LSVC: An Example
  - Large-Scale Visual Surveillance
- Conclusions

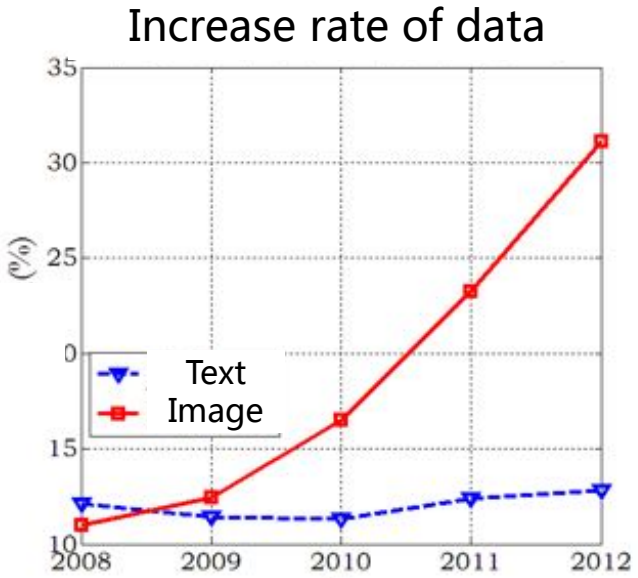
Large Scale Visual Computing (LSVC): Definition

- **Visual Computing**
  - Processing and analyzing visual information
- **Large Scale Visual Computing**
  - Processing and analyzing "large" scale visual information
    - Large scale in volume
    - Large variety in categories
    - Large inhomogeneity in properties

LSVC: Importance

Exponential Growth in Visual Data





Practical and urgent need for LSVC

Proliferation of Surveillance Cameras

India's UID Program

AADHAAR

LSVC : Major Challenges (1)

- **Large scale in volume:**
  - **Computational efficiency**

50 hours new videos every minutes

300 million new photos everyday

LSVC : Major Challenges (2)

- **Large variety in categories:**
  - **Generalizability and applicability**

VOC dataset：

20 object classes, <20K images

Top 1 Recognition Accuracy：>90%

ImageNet dataset：

1000 object classes, >1.3m images

Top 1 Recognition Accuracy：< 70%

LSVC : Major Challenges (3)

- **Large inhomogeneity in properties:**
  - **Robustness and stability**

documents, images, videos, voice…

LSVC : Key Problems

NIPS2012 Workshop :
Large Scale Visual Recognition and Retrieval

Main Topics

- **Feature representations**
- **Model learning**
- **Transfer learning**
- **Datasets issues**
- **……**

Outline

- Background and Context
- **Feature Representation**
- Model Learning
- LSVC: An Example
  - Large-Scale Visual Surveillance
- Conclusions

Feature Representation

documents, images, videos, voice…

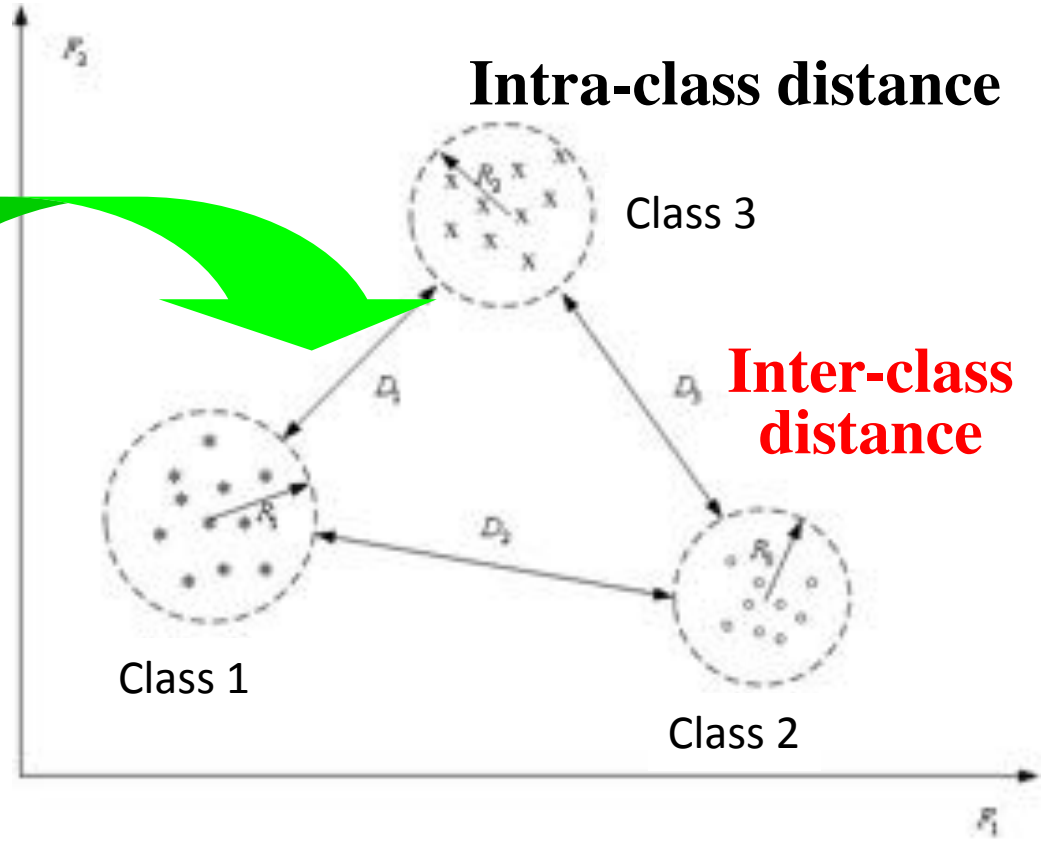
Multi-modality

Multi-source

How to find robust representation for multi-modality and multi-source big visual data?

Robust feature representation

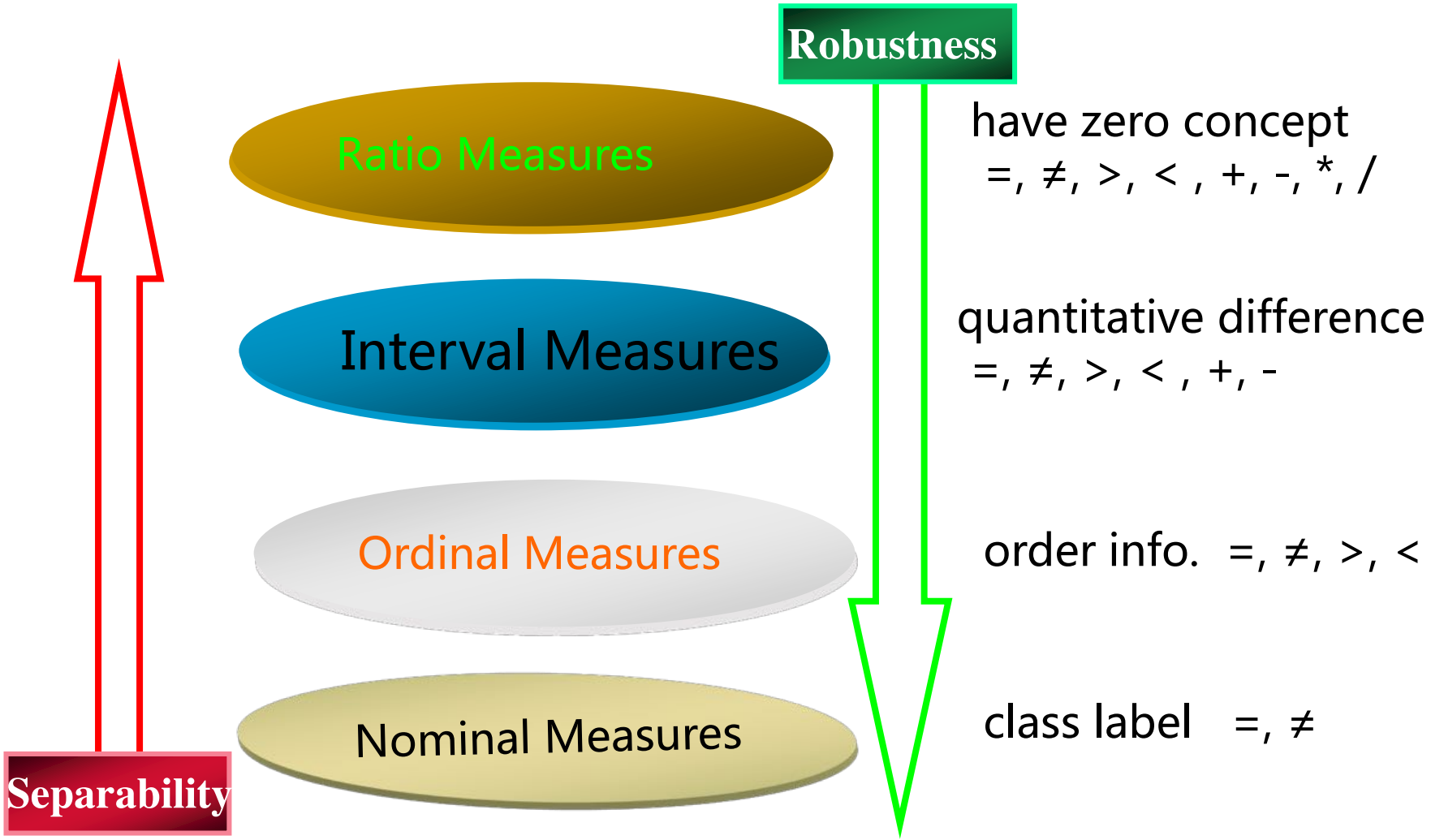
Feature Representation: Mapping from Image Space to Feature Space



Maximize inter-class distance D → **Separability**

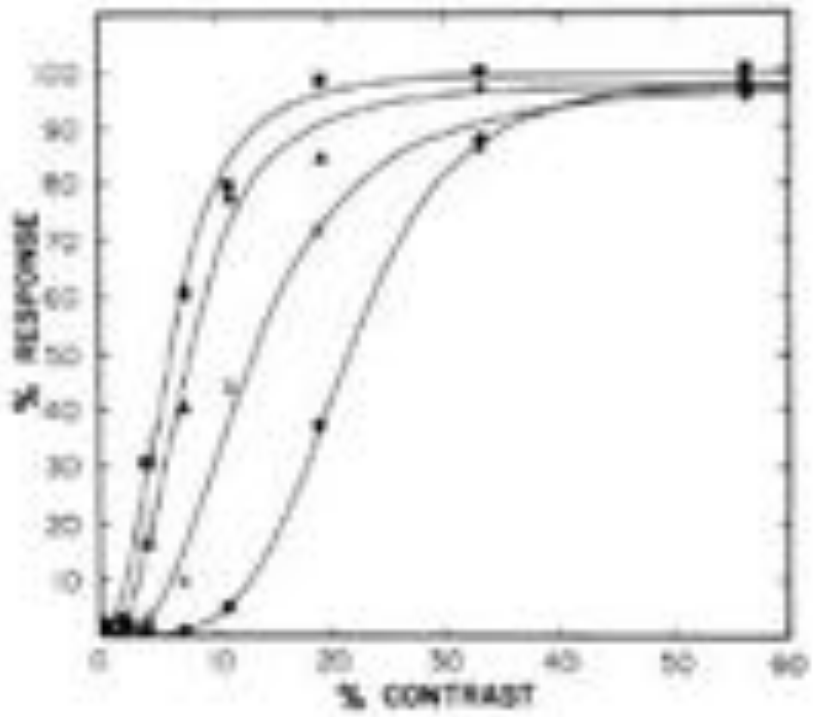
Minimize intra-class distance R → **Robustness**

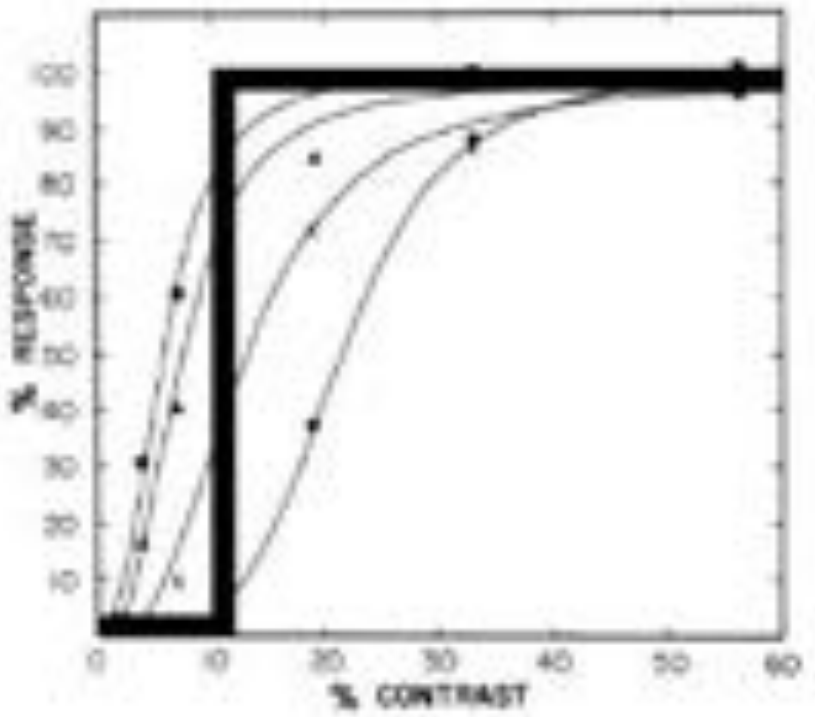
Ordinal Feature Representation



S. Stevens, "On the Theory of Scales of Measurement", Science, Vol.103, No.2684, pp.677-680, June 1946.

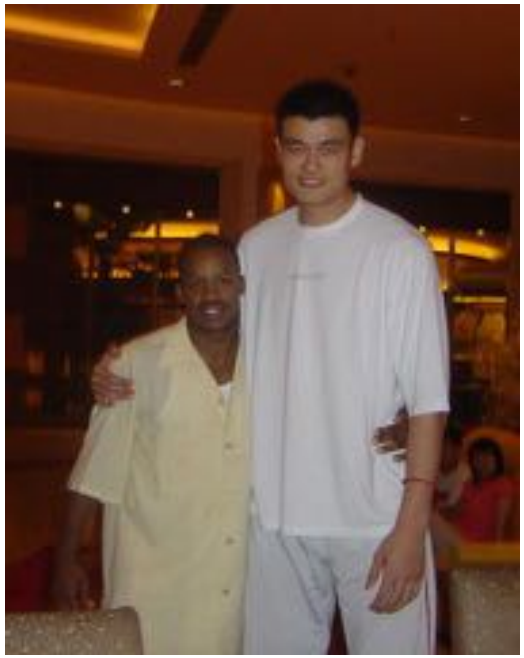
Biological Support for OM





Duane G. Albrecht and David B. Hamilton. Striate cortex of the monkey and cat: Contrast response function. *Journal of Neuroscience*, 48(1):217–237, July 1982.

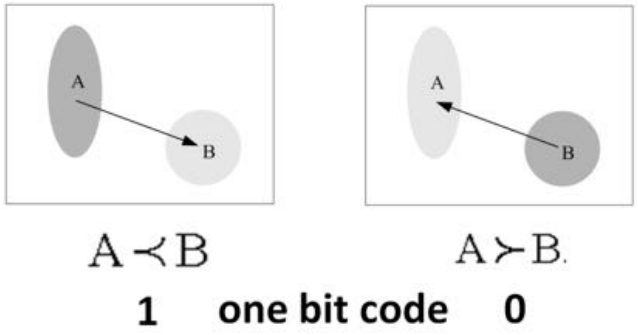
Ordinal Measures (OM) in Everyday Life

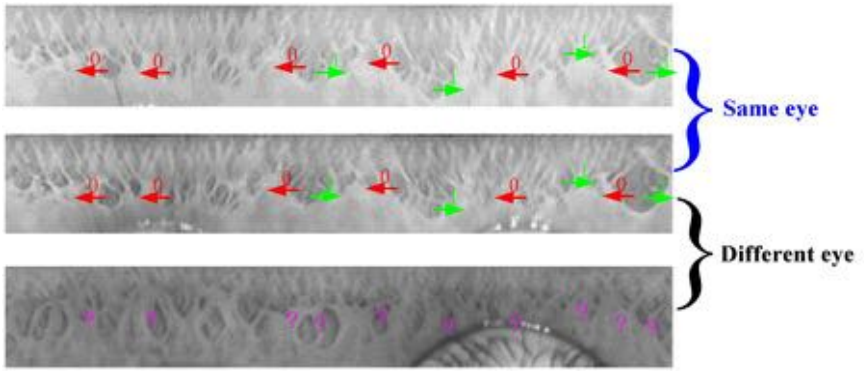


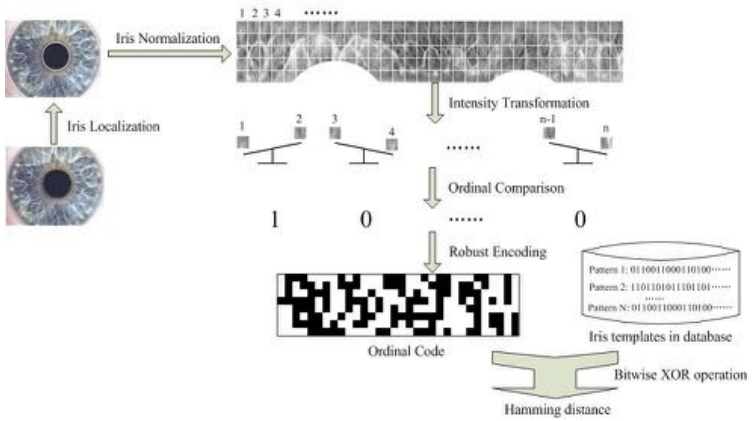
Height

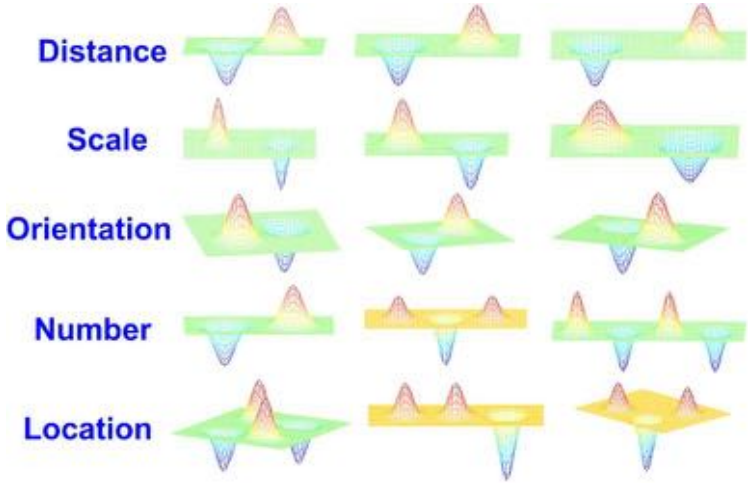
Weight

We have developed very efficient and robust iris recognition algorithms based on ordinal measures.









Zhenan Sun and Tieniu Tan, "Ordinal Measures for Iris Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 12, 2009, pp. 2211 - 2226.

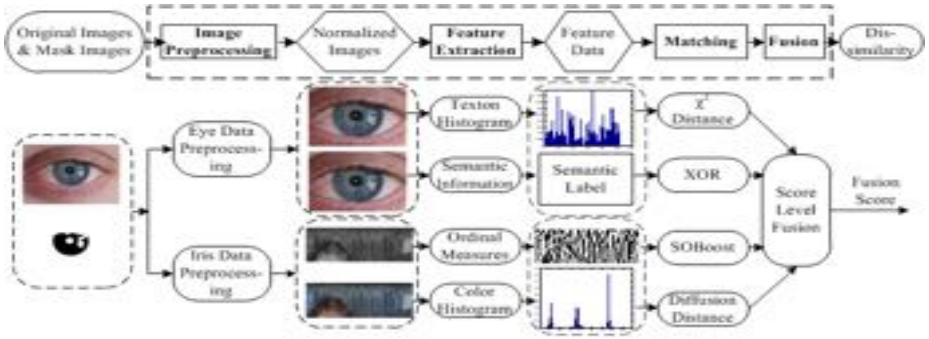
NICE:II

Noisy Iris Challenge Evaluation - Part II

!--[if !supportLists]--> <!--[endif]-->**September 06th, 2010:** The classification of the NICE.II best participants is available on the contest website. Additionally, the full classification results were sent by email to all participants.

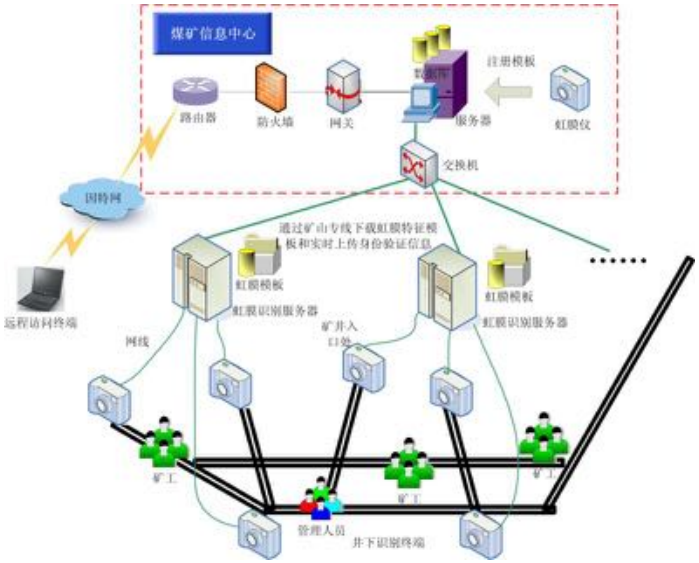
Participants invited to publish their approach in the Pattern Recognition Letters Journal:

Ranking	Unsername	Affiliation	Country	Decidability (d')
1	CASIA	National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences	China	2,5748
2	Betaeye	Techshino Biometrics Research Center,Department of Mathematics, Northeastern University	China	1,8213
3	UBI	University of Beira Interior	Portugal	1,7786
4	Parkgr	Biometrics Engineering Research Center (BERC), Dongguk University	Republic of Korea	1,6398
5	PeihuaLi	College of Computer Science and Technology, Heilongjiang University	China	1,4758
6	BIPLab	University of Salerno	Italy	1,2565
7	HLJUCS	College of Computer Science and Technology, Heilongjiang University	China	1,1892
8	DMCS	Technical University of Lodz	Poland	1,0931



Tieniu Tan, Xiaobo Zhang, Zhenan Sun, and Man Zhang, "Noisy Iris Image Matching by Using Multiple Cues", Pattern Recognition Letters, 2012. (Invited paper)

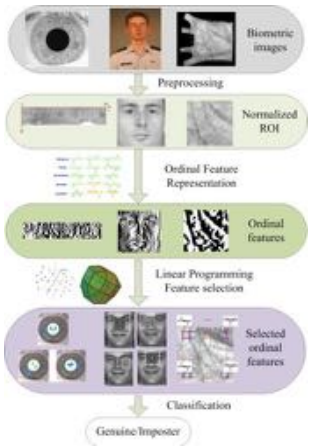
Iris Recognition for Coal Miner Identification

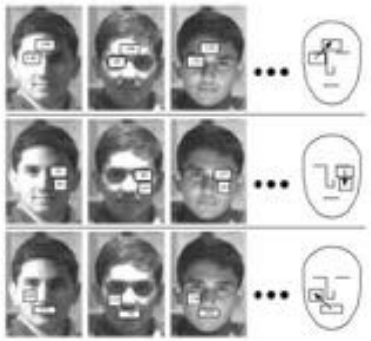


http://www.IrisKing.com

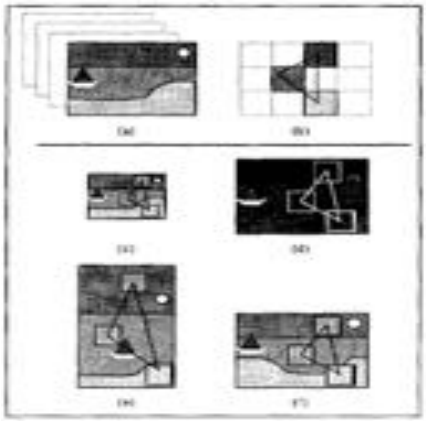
Iris Recognition at A Distance

Other Applications of Ordinal Measures





Biometrics



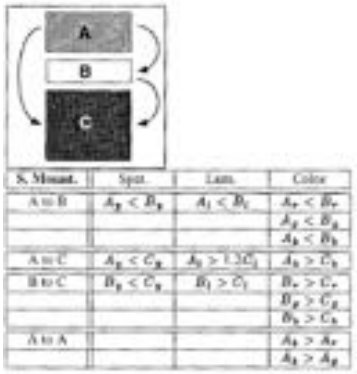
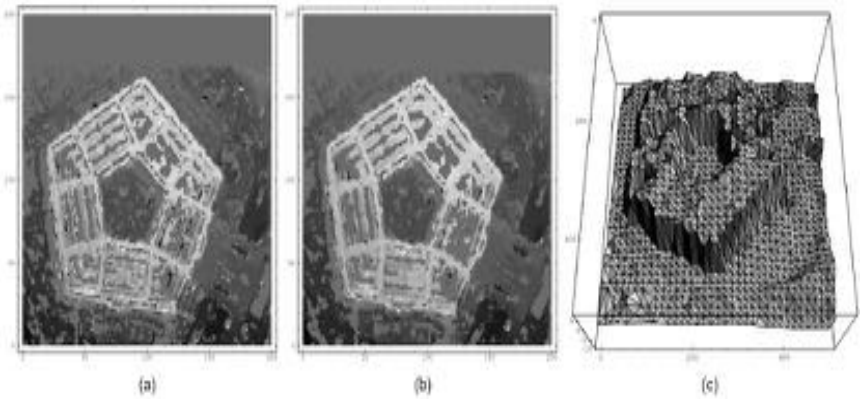


Image retrieval



Stereo vision

Object detection

Outline

- Background and Context
- Feature Representation
- **Model Learning**
- LSVC: An Example
  - Large-Scale Visual Surveillance
- Conclusions

Model Learning

Large inhomogeneity in properties

ImageNet：1000 categories

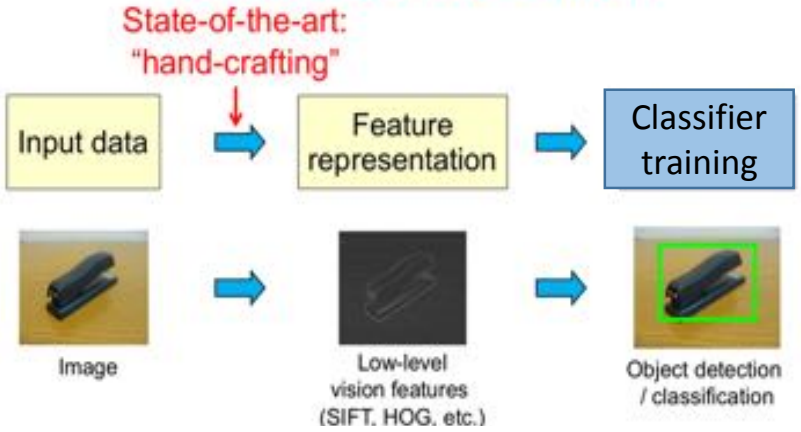
Large variety in categories

Manually designed features are often unable to cope with large inhomogeneity in properties and large variety in categories

Directly learn models from data

A Breakthrough: Deep Learning

Traditional pipeline



Need rich experience for hand-crafting feature design

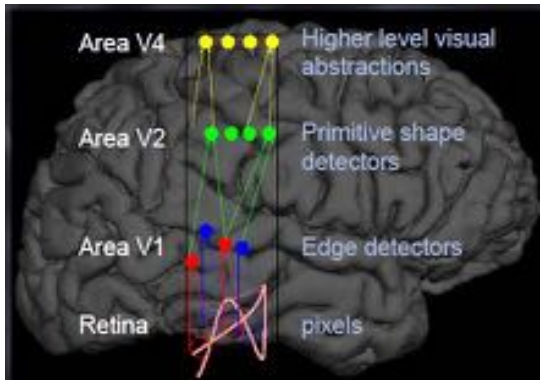
End-to-end Model Learning

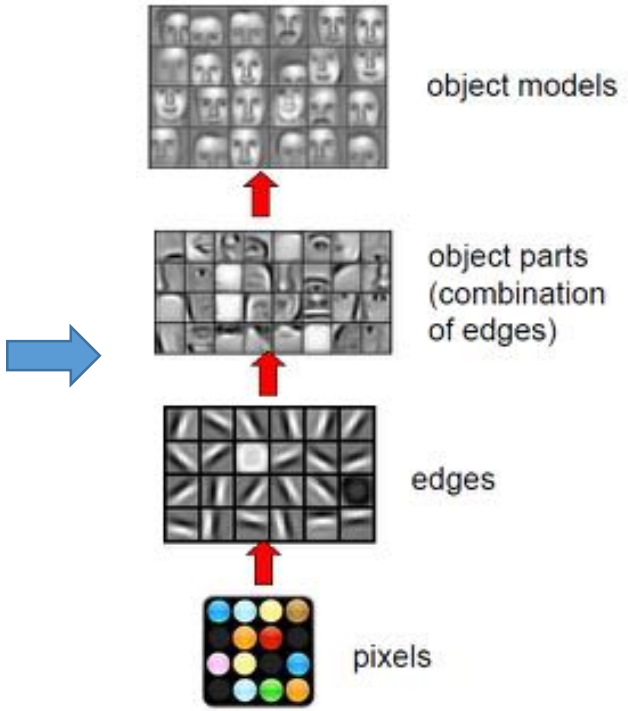
- **No hand-crafting feature design**
- **Directly map from pixels to labels**
- **Use big visual data to deal with varieties**

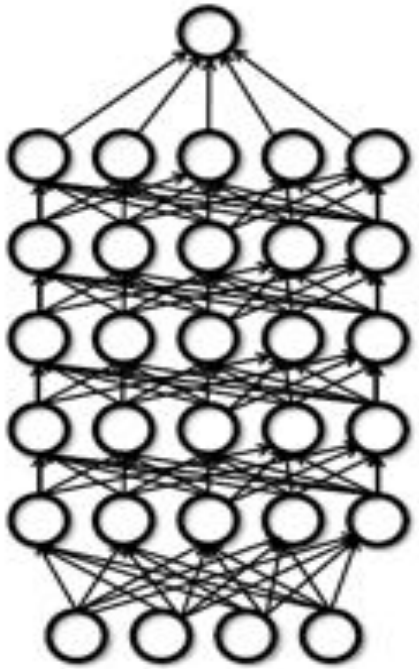
The answer is hidden in the big data!

An important motivation of deep learning is the big visual data

Deep Learning + X

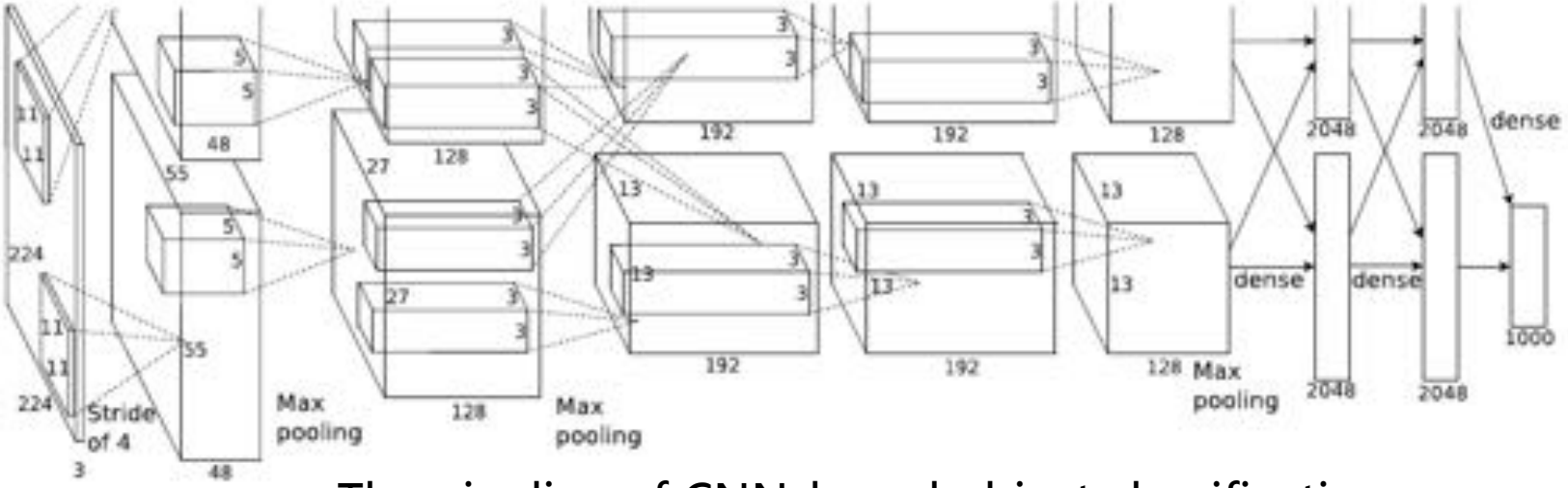






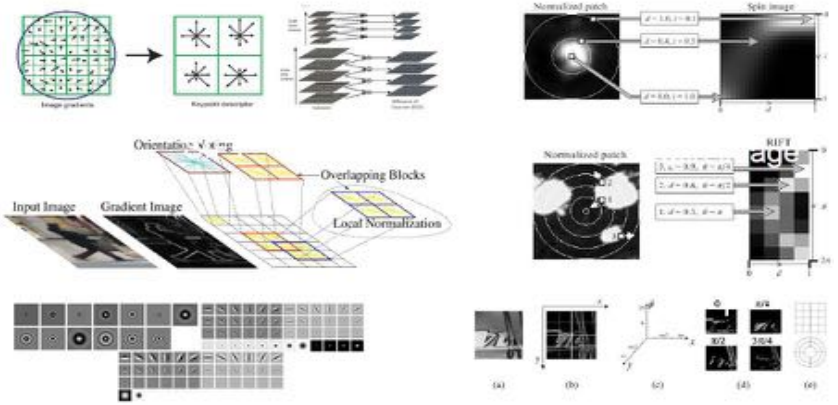
Biologically inspired deep learning promotes the fast development of many fields in computer vision

Object Classification



The pipeline of CNN-based object classification

Visualization of the first layer



Traditional hand-crafting features

A. Krizhevsky and J. Hinton, Image Classification Based on Convolutional Neuron Networks, NIPS2012

Object Classification on ImageNet2012

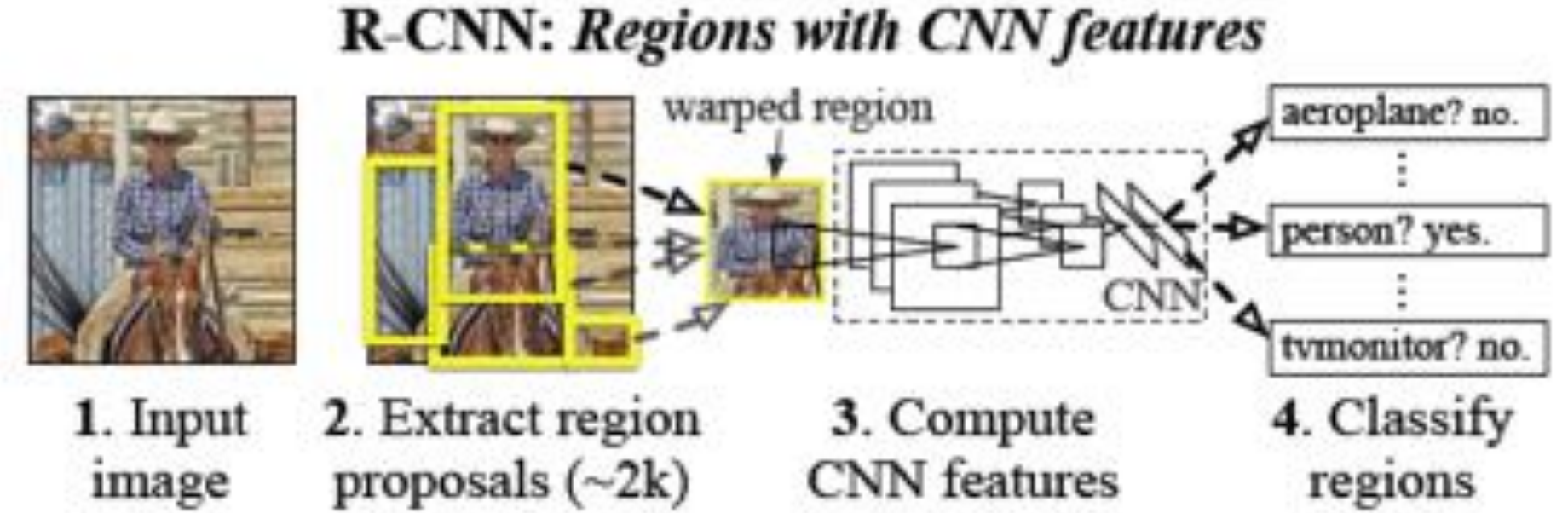
Traditional shallow models
72% 2010
74% 2011

CNN based deep models
85% 2012
89% 2013
94% 2014
>95% 2015

Human: 95%

http://imagenet.org/challenges/LSVRC/{2010,2011,2012,2013,2014}

Object Detection

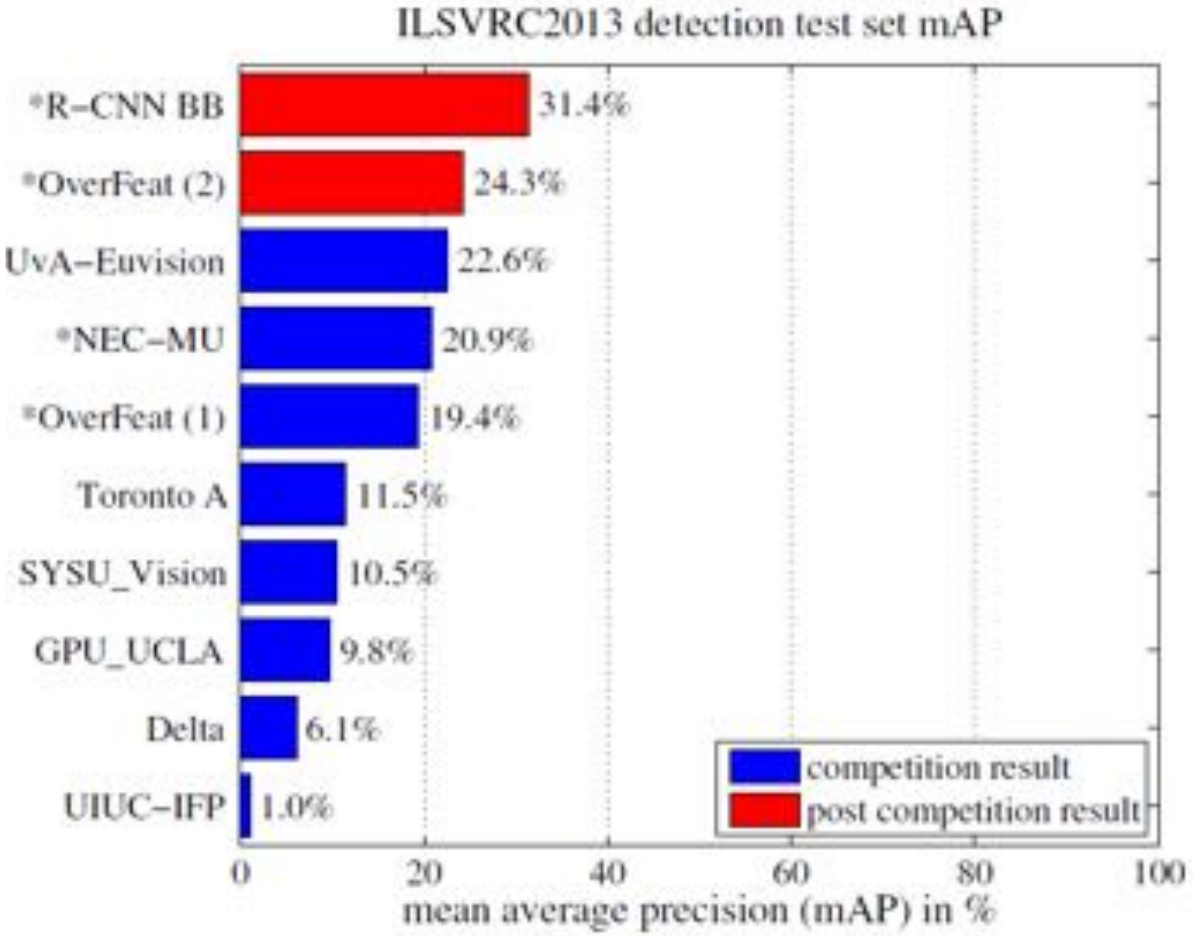


VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [20]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [39]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Regionlets [41]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [18]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

Some best results on PASCAL VOC2010

Ross et al., Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR2014

200 categories object detection competition on ImageNet2013

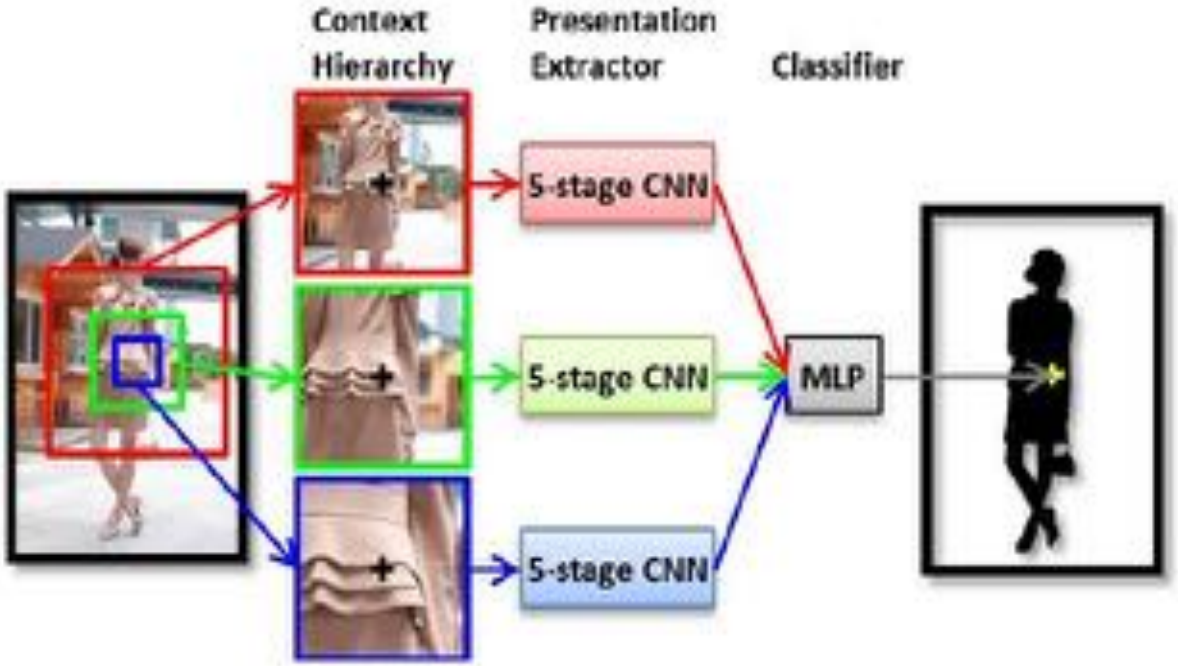


The previous best traditional algorithm (deformable part model) can only achieve **9%** in accuracy

Now the best result (based on deep learning) is more than **50%** in accuracy.
W. Ouyang, et. al, DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection, CVPR2015

Object Segmentation

Multi-scale CNN for human segmentation



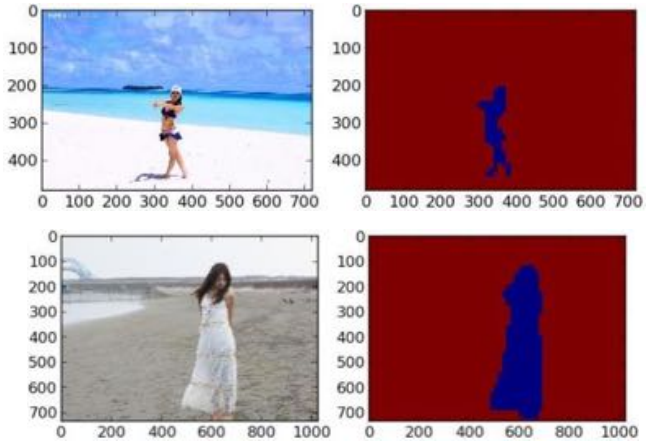
Novelty:

- **For each pixel, we use three different scales of windows, each of which is described by a powerful CNN to model the relationship between a pixel and its thousands of neighboring pixels.**
- **Multi-scale CNNs are robust to different scales of object segmentation**

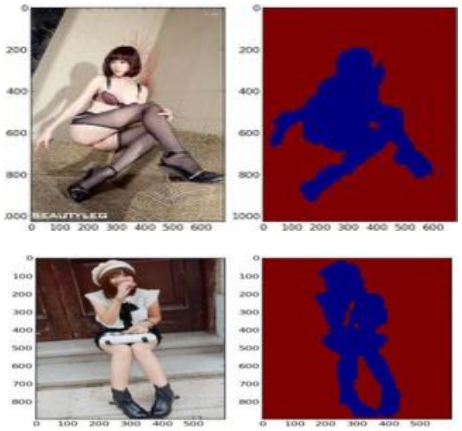
Champion in a human segmentation competition

Rank	Name	Score	Publish Time
1	NLPR&CRIPAC	0.8683	Tue Oct 15 CST 2013
2	Freedom	0.7817	Tue Oct 15 CST 2013
3	WEESEE	0.7600	Tue Oct 15 CST 2013
4	CASIA_IGIT	0.7595	Tue Oct 15 CST 2013
5	DT_ppseg	0.7587	Tue Oct 15 CST 2013
6	FlyHigh	0.7328	Tue Oct 15 CST 2013
7	EagleEye	0.7213	Tue Oct 15 CST 2013
8	sysu_vision	0.7167	Tue Oct 15 CST 2013
9	DeepLearner	0.6111	Tue Oct 15 CST 2013
10	RandomForest	0.5117	Tue Oct 15 CST 2013

Complex backgrounds



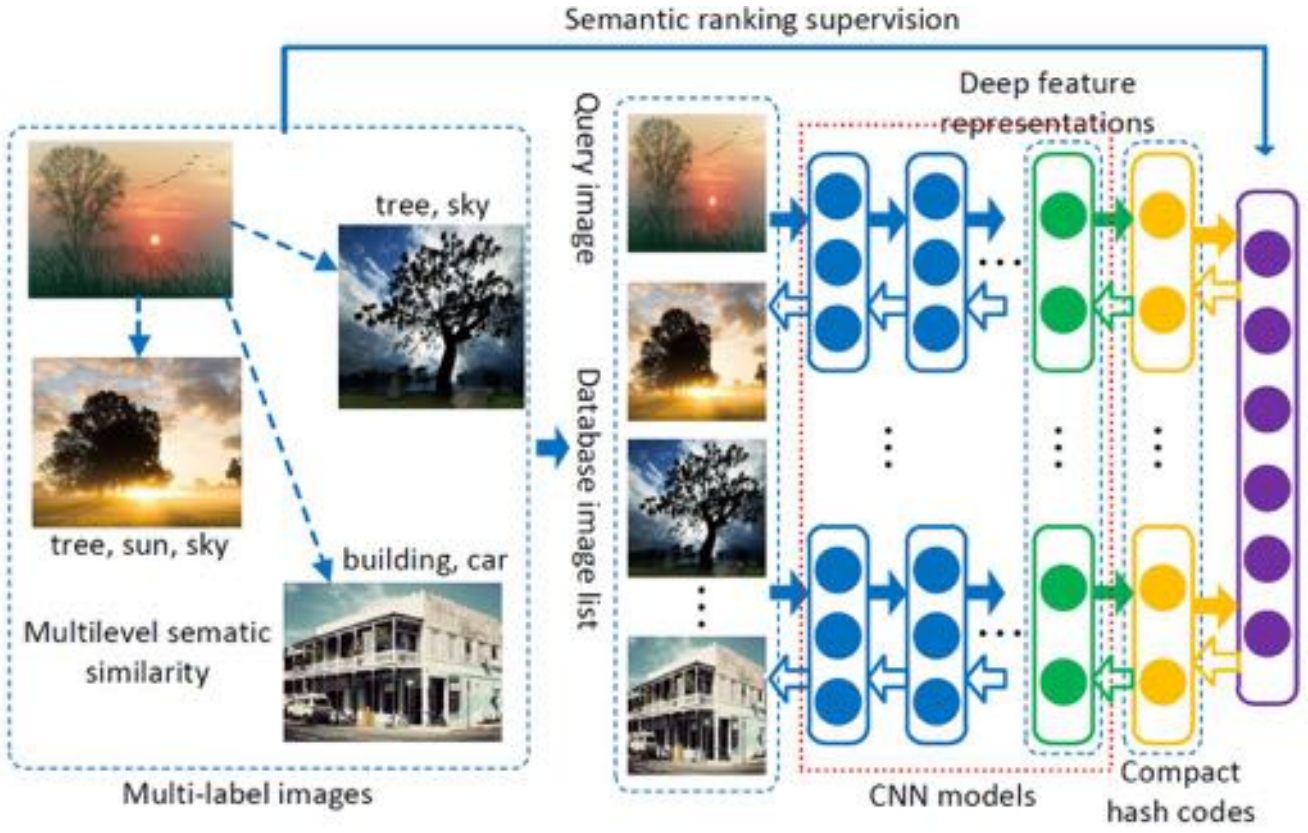
Multiple scales



Various poses

Input Image | Manual Results | Our Results

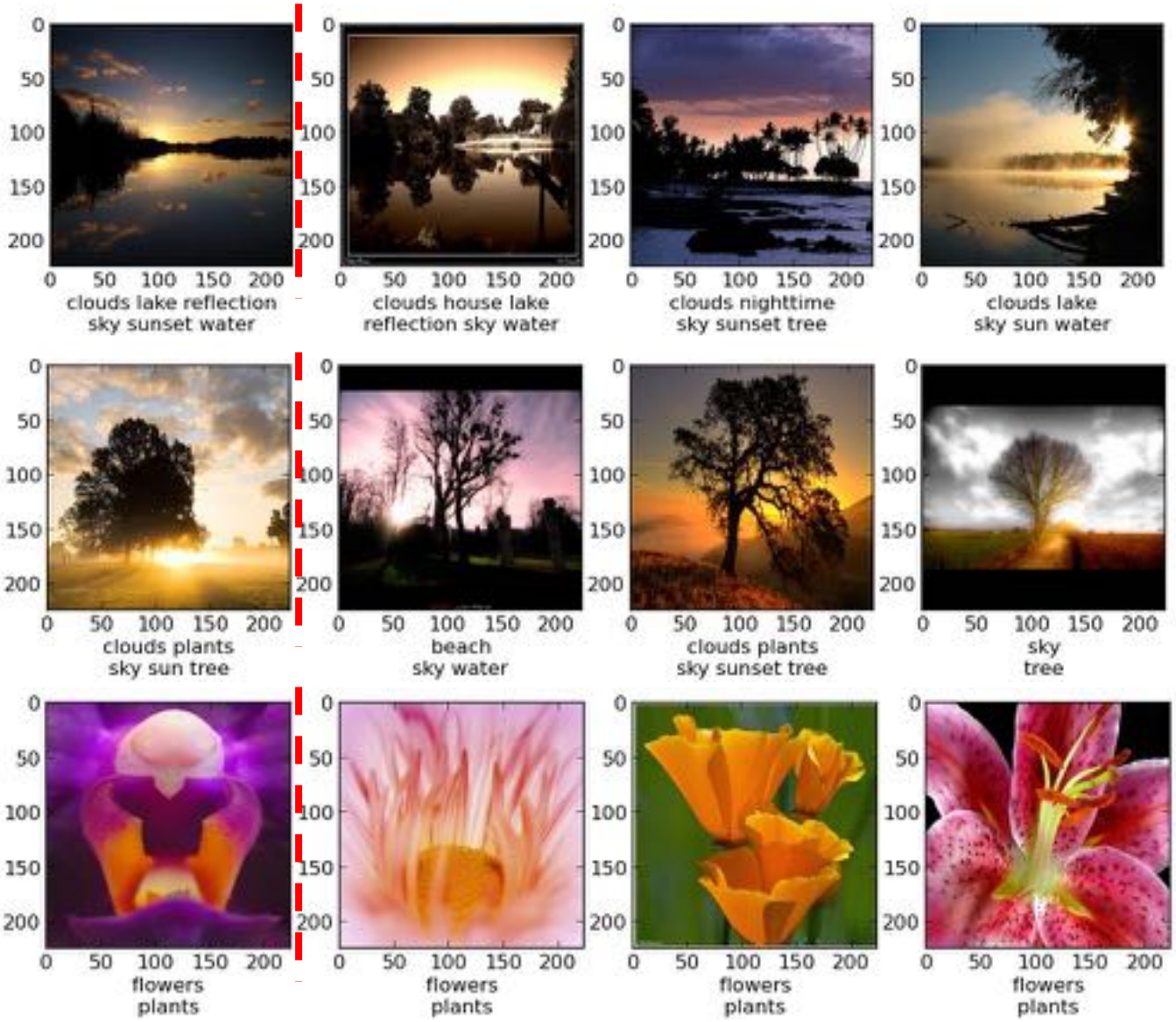
Object Retrieval



Novelty: Combine semantic ranking and CNN to address the problem of preserving multilevel semantic similarity for hashing

Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan, Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval, CVPR2015

Example results of our retrieval algorithm



Example results of our retrieval algorithm

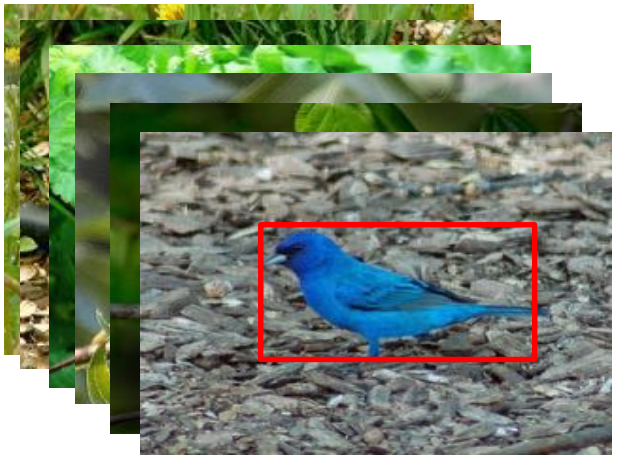
Deep Learning: A Panacea?

- ***Deep*** **learning is a** ***shallow*** **model of brain mechanisms**
- **When is it deep enough or how deep is deep?**
- **Theoretical explanation of its effectiveness?**

- **Computational complexity**
- **Access to large amount of annotated or labeled data for training**

In large-scale applications, annotating object location can be very laborious and can also be ambiguous.

Strong Annotation

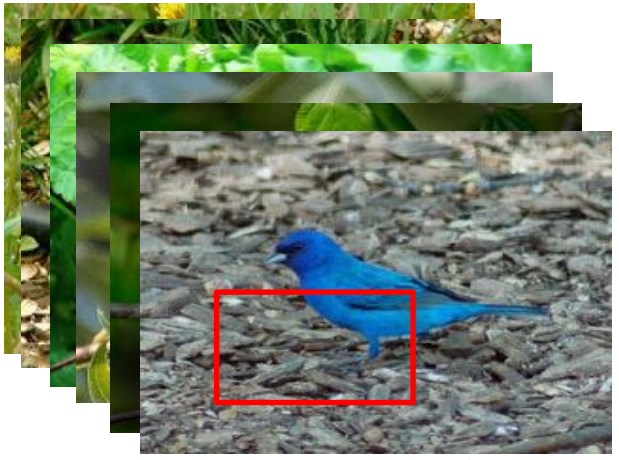


Category: Yes
Location: Yes

Weak Annotation

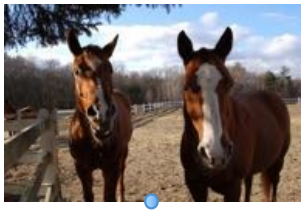
Category: Yes Location: No

Noisy Annotation



Category: Yes Location: Noisy

Weakly Supervised Object Detection



Positive Images

Negative Images

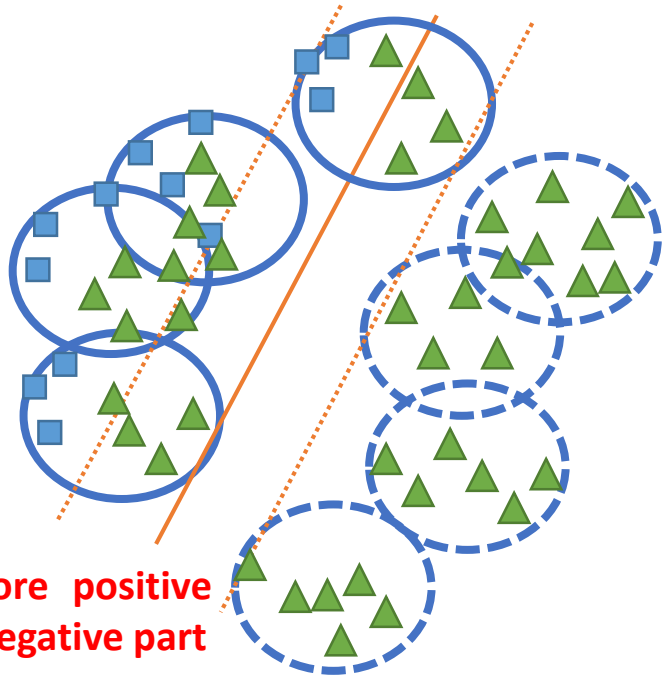
Input Images

Region Proposals

Novelty: To reduce ambiguity in positive images, we present a bag-splitting algorithm that iteratively generates new negative bags from positive ones

Conv. Nets

Representations



MILinear Mining

Comparison of detection results (Average Precision, AP) on VOC2007 dataset. The first 3 rows are the state-of-the-art results for supervised object detection. The last 7 rows list the results from weakly supervised localization.

Methods	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM-v4 [29]	29.0	54.6	0.6	13.4	26.2	39.4	46.4	16.1	16.3	16.5	24.5	5.0	43.6	37.8	35.0	8.8	17.3	21.6	34.0	39.0	26.3
DPM-v5 [42]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
R-CNN fc6 [33]	56.1	58.8	34.4	29.6	22.6	50.4	58.0	52.5	18.3	40.1	41.3	46.8	49.5	53.5	39.7	23.0	46.4	36.4	50.8	59.0	43.4
Siva *et al.* [12]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
OCP[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15.0
Cover+SVM [18]	23.4	43.5	22.4	8.1	6.2	33.9	33.8	30.4	0.1	17.9	11.5	17.1	24.7	40.2	2.4	14.8	21.4	15.1	31.9	6.2	20.3
Cover+LSVM [18]	28.2	47.2	17.6	9.6	6.5	34.7	35.5	31.5	0.3	21.7	13.2	20.7	25.2	39.8	12.6	18.6	21.2	18.6	31.7	10.2	22.2
Cover+SLSVM [18]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Multi-fold MIL [17]	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23	4.9	14.1	31.9	41.9	19.3	11.1	27.6	12.1	31	40.6	22.4
Ours(noBS)	40.9	38.3	22.0	5.7	1.6	37.9	44.7	15.2	0.8	33.3	13.2	20.0	28.7	42.6	19.1	17.0	20.4	20.9	34.3	11.3	23.4
Ours(BS1)	41.3	39.7	22.1	9.5	3.9	41.0	45.0	19.1	1.0	34.0	16.0	21.3	32.5	43.4	21.9	19.7	21.5	22.3	36.0	18.0	25.4

Weiqiang Ren, Kaiqi Huang, Dacheng Tao, Tieniu Tan. Weakly Supervised Large Scale Object Localization with Multiple Instance Learning and Bag Splitting, IEEE TPAMI (to appear)


plane
bike
bird
boat
bottle
bus
car
cat
chair
cow
table
dog
horse
bike
person
plant
sheep
sofa
train
tv

Outline

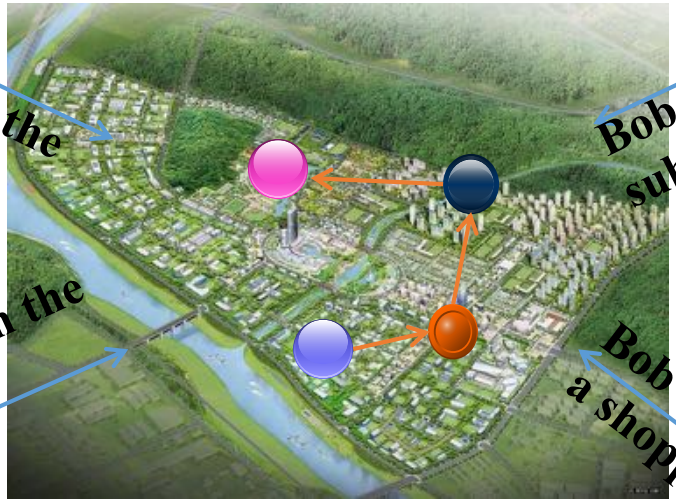
- Background and Context
- Feature Representation
- Model Learning
- **LSVC: An Example**
  - **Large-Scale Visual Surveillance**
- Conclusions

Large-Scale Visual Surveillance

Large-Scale Visual Surveillance (LSVS)

Bob is in the airport.

Bob is on the subway.

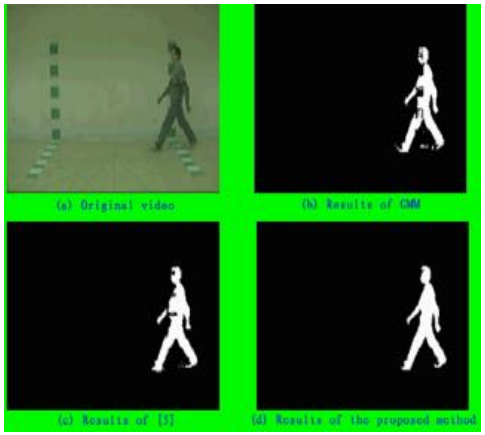


Bob is in the CBD.

Bob walks down a shopping street.

Key Problems in LSVS

- **Background modeling**



- **Object detection**

- **Object tracking**

- **Behavior recognition**

LSVS: Major Challenges

- **Multi-modality/multi-distribution**
  - Multi-view, occlusion, multi-scale, cluttered background, etc.

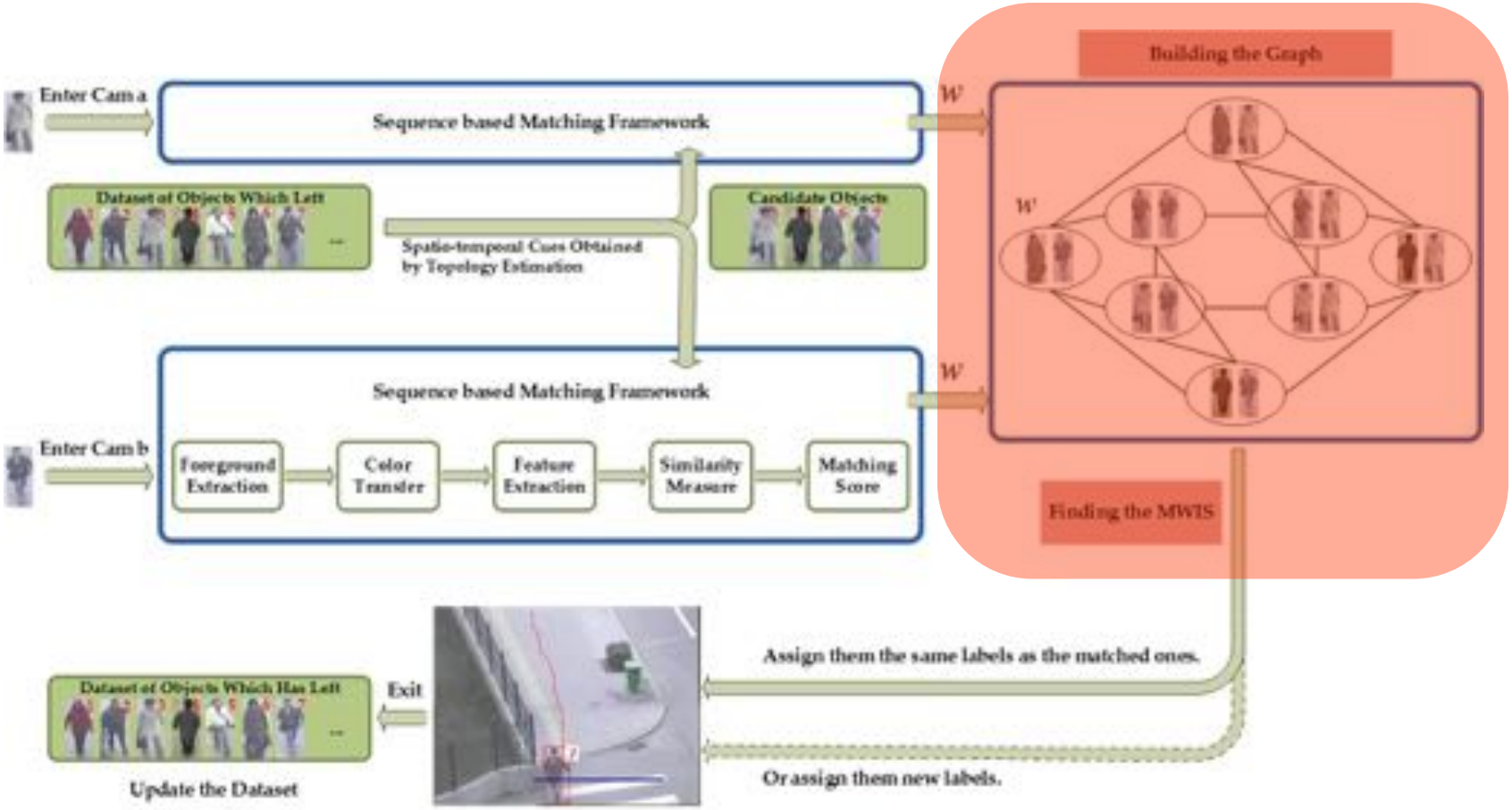
- **Model scalability**
  - How to scale up to large scale applications: scalable feature representation, scalable model inference, etc.

Multi-object Tracking in Non-overlapping Camera Networks

Problem Statement:

- **Establish object correspondences between non-overlapping cameras to make sure each object has a unique identity across the entire camera network.**

Multi-object Tracking in Non-overlapping Camera Networks

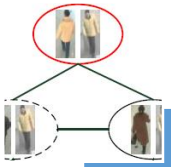


The problem is formulated as an integer programming problem solved by finding the maximum weight independent set (MWIS) in a global graph.

Constructing the Graph

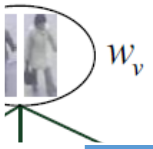
Nodes

- Nodes are candidate matches (a departure object vs. an arrival object).
- Depend on topology estimation to get inter-camera spatio-temporal cues.



Edges

- Two nodes are linked together when sharing common objects, i.e., conflicting with the fact that an object can not appear in two different places at the same time.

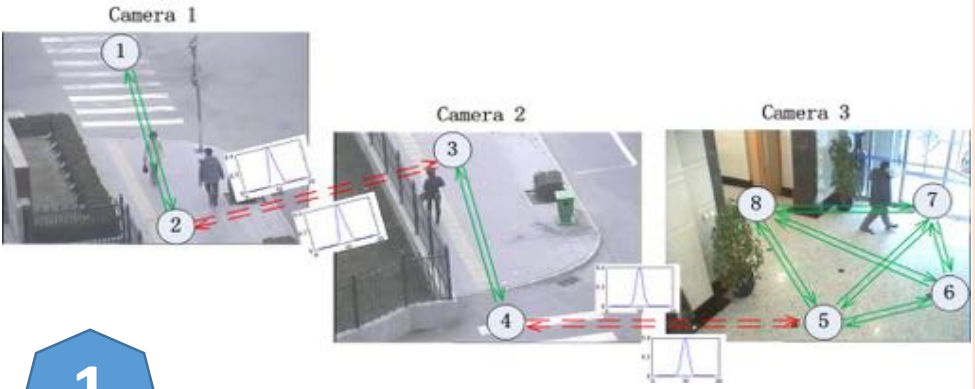


Weights

- Weights are similarities based on object matching.
- Need color transfer to deal with illumination variance between different views.

Key Problems

- **Topology estimation**



1

- **Color transfer between cameras**

2

- **Object matching**

3

- **Find the MWIS solution**
  - NP-hard

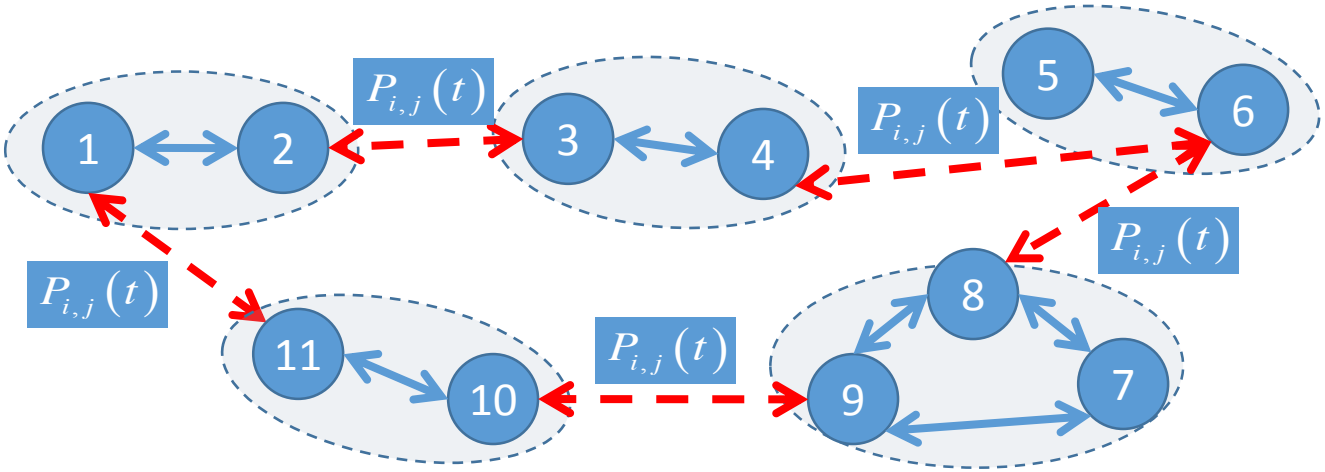
$$IP: \quad \max \sum_{v=1}^{N_V} w_v \chi_v$$

$$s.t. \ \forall v \in V, \quad \chi_v \in \{0,1\}$$

$$\forall (v,v') \in E, \quad \chi_v + \chi_{v'} \leq 1$$

4

1 Topology Estimation



- Nodes, defined either as the entry/exit zones in the field of views of cameras or single cameras.
- Links across cameras, which indicate the connectivity of nodes.
- Transition time probability distribution for each link, demonstrating the average transition time of objects moving from one node to another.

Topology Estimation based on N-neighbor Accumulated Cross-correlation Functions

- **Propose N-neighbor accumulated cross-correlation functions based on the following observation:**
  - Transition time by correct matches is generally distributed around the average transition time, while transition time by wrong matches is randomly distributed.

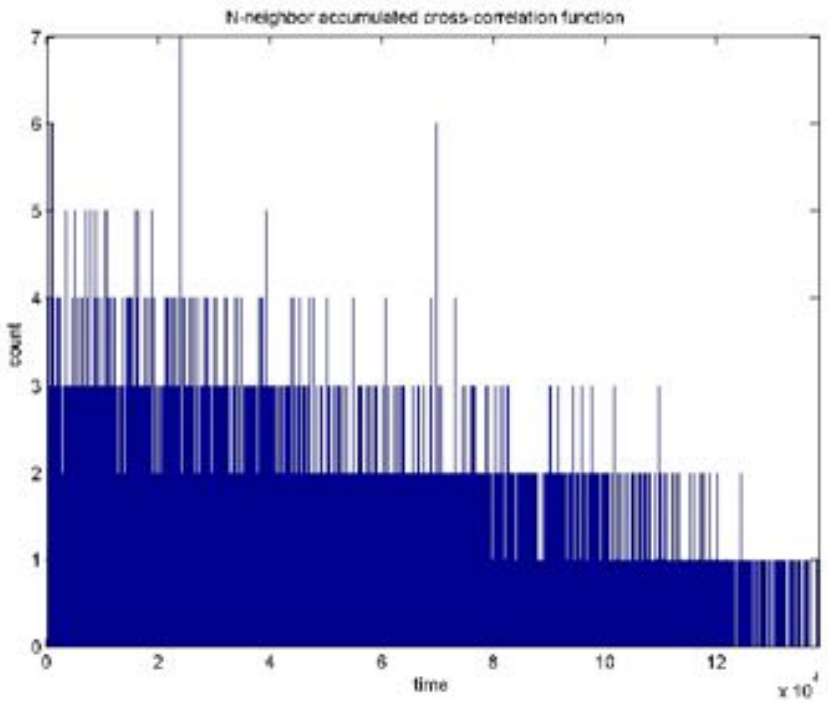
$$R_n^{i,j}(\tau_n) = \sum_{\tau_0=\tau_n-n}^{\tau_n+n} R_0^{i,j}(\tau_0)$$
$$= \sum_{\tau_0=\tau_n-n}^{\tau_n+n} E\left[D_i(t)\cdot A_j(t+\tau_0)\right]$$
$$= \sum_{\tau_0=\tau_n-n}^{\tau_n+n}\sum_{t=-\infty}^{+\infty} D_i(t)\cdot A_j(t+\tau_0),\ \tau_n \geq n$$

$D_i(t)$ is the departure time sequence observed at node i, and $A_j(t)$ is the arrival time sequence observed at node j.

Unlike previous cross-correlation based work, our topology recovering method can deal with large amounts of data without considering the size of time window.

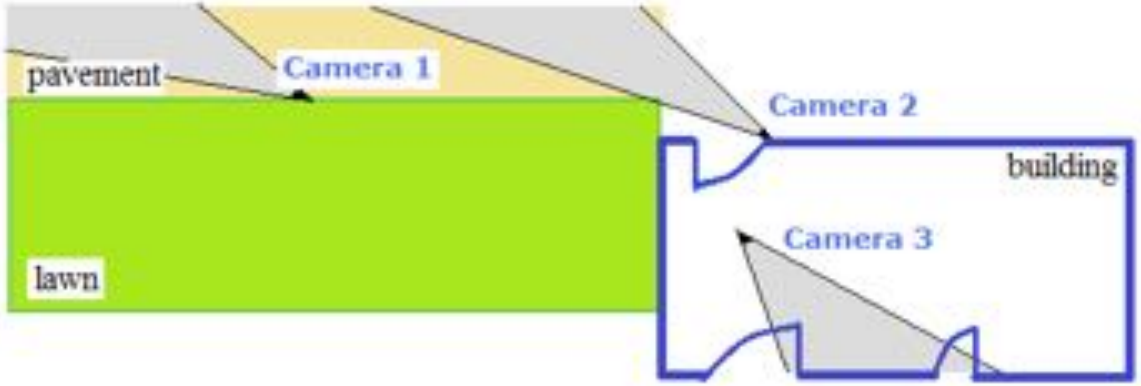
Topology Estimation based on N-neighbor Accumulated Cross-correlation Functions

- Generally, N-neighbor accumulated cross-correlation function converges quickly when there is a link between these two nodes (entry/exit zones).
- The location of the clear peak denotes average transition time between these two nodes.

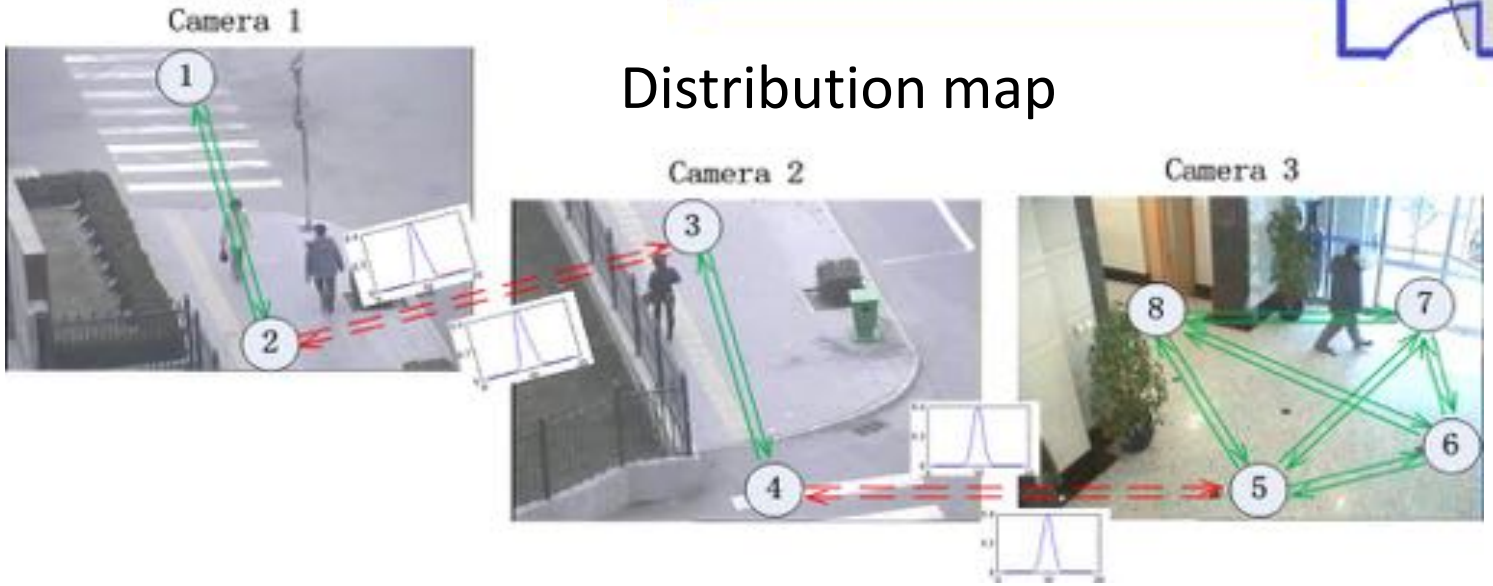


Experimental Results

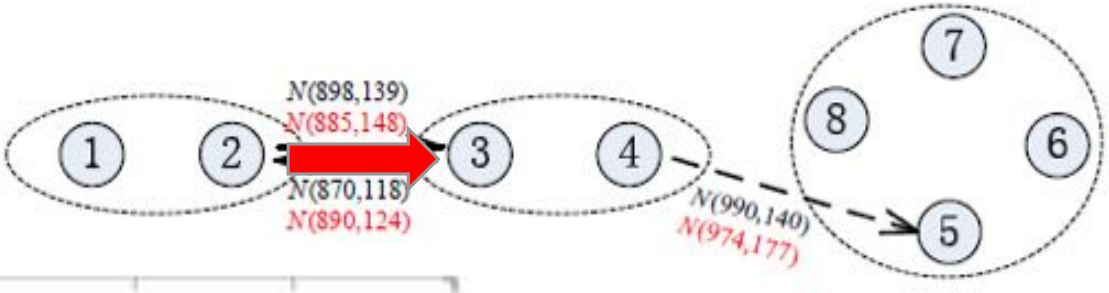
- Real-world three-camera network



Distribution map

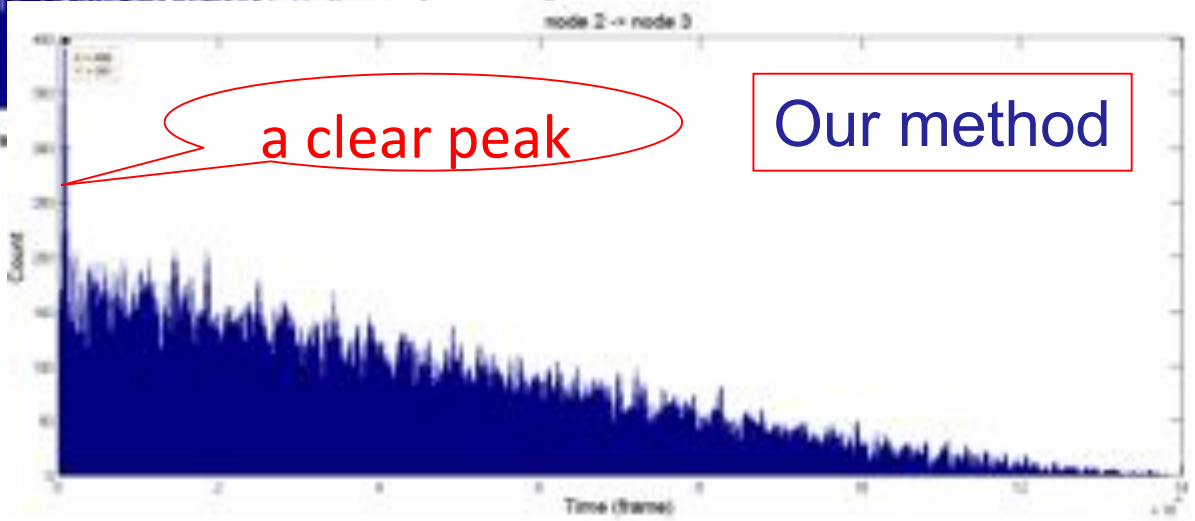


Experimental Results



Topology

[Makris 2004]



Our method

X.T. Chen, K.Q. Huang and T.N. Tan, Learning the three factors of a non-overlapping multi-camera network topology, CCPR, 2012. (Best Paper)

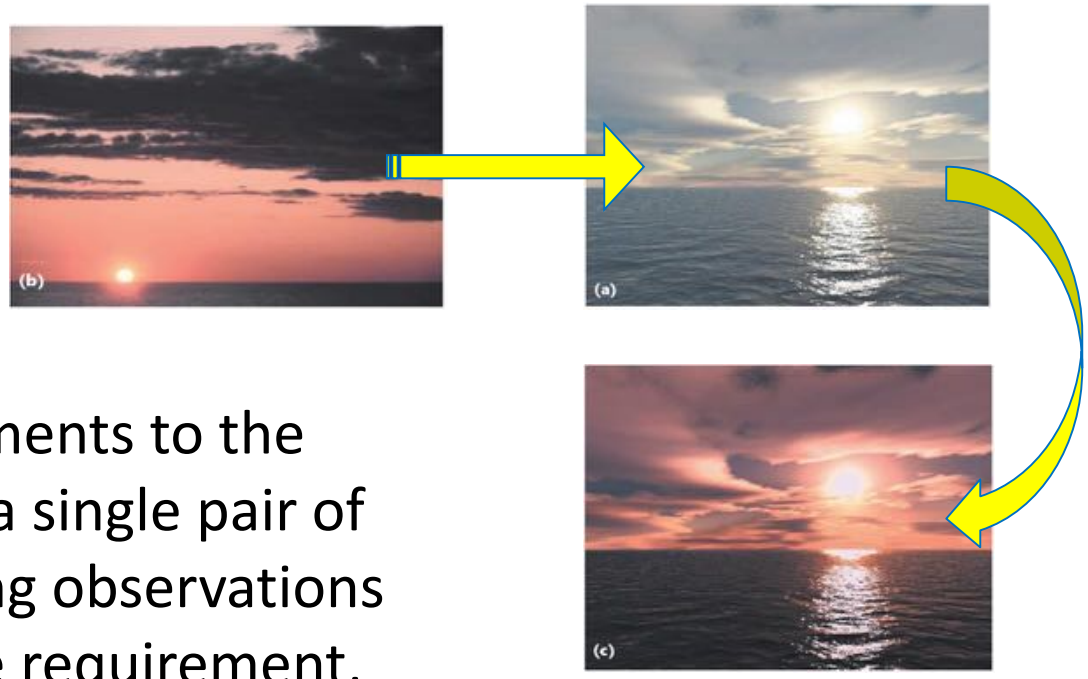
2 Color Transfer between Cameras

- Problem
  - Illumination changes with both the viewpoints (cameras) and time.

  - Illumination variance has a large influence over the appearance of objects.

Color Transfer between Cameras

- Our method
  - Apply a color characteristic transfer method to impose the color characteristics of a target image on a source image [Reinhard 2001]



Advantage:

- Low requirements to the training set: a single pair of corresponding observations can meet the requirement, making updates efficient.

Algorithm Description

- Step 1：given a target-source image pair, calculate the corresponding color characteristic transfer (CCT) model：

$$\left\{\frac{\sigma_t^l}{\sigma_s^l}, \frac{\sigma_t^\alpha}{\sigma_s^\alpha}, \frac{\sigma_t^\beta}{\sigma_s^\beta}, m_s^l, m_s^\alpha, m_s^\beta, m_t^l, m_t^\alpha, m_t^\beta\right\}$$

Target image

Source image

Transferred image

Algorithm Description

- Step 2: the source image is transferred to the target image appearance using the CCT model .

Under Cam_x

$$l_y^a = \frac{\sigma_t^l}{\sigma_s^l}(l_x^a - m_s^l) + m_t^l$$

$$\alpha_y^a = \frac{\sigma_t^\alpha}{\sigma_s^\alpha}(\alpha_x^a - m_s^\alpha) + m_t^\alpha$$

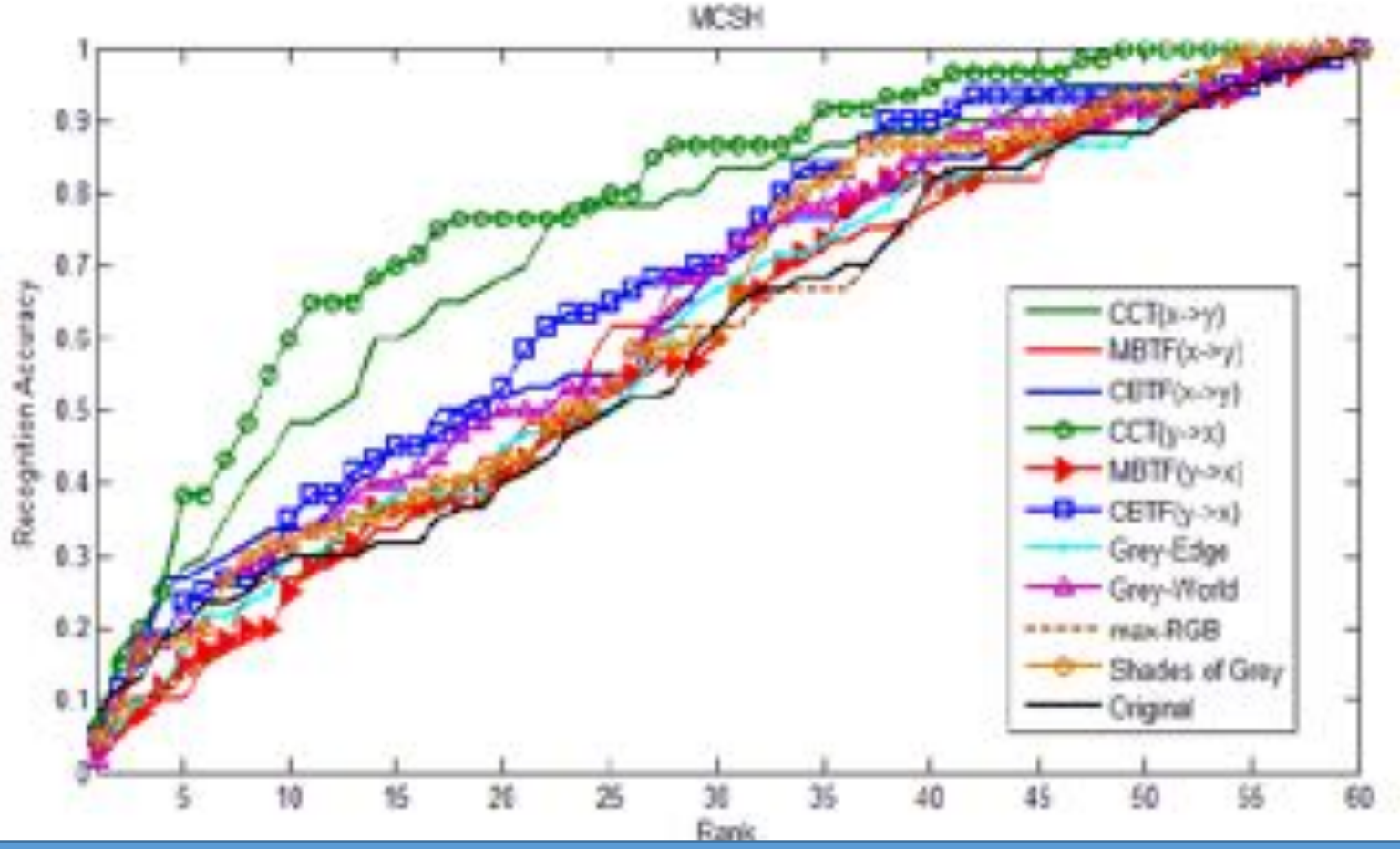
$$\beta_y^a = \frac{\sigma_t^\beta}{\sigma_s^\beta}(\beta_x^a - m_s^\beta) + m_t^\beta$$

Estimated appearance under Cam_y

VS

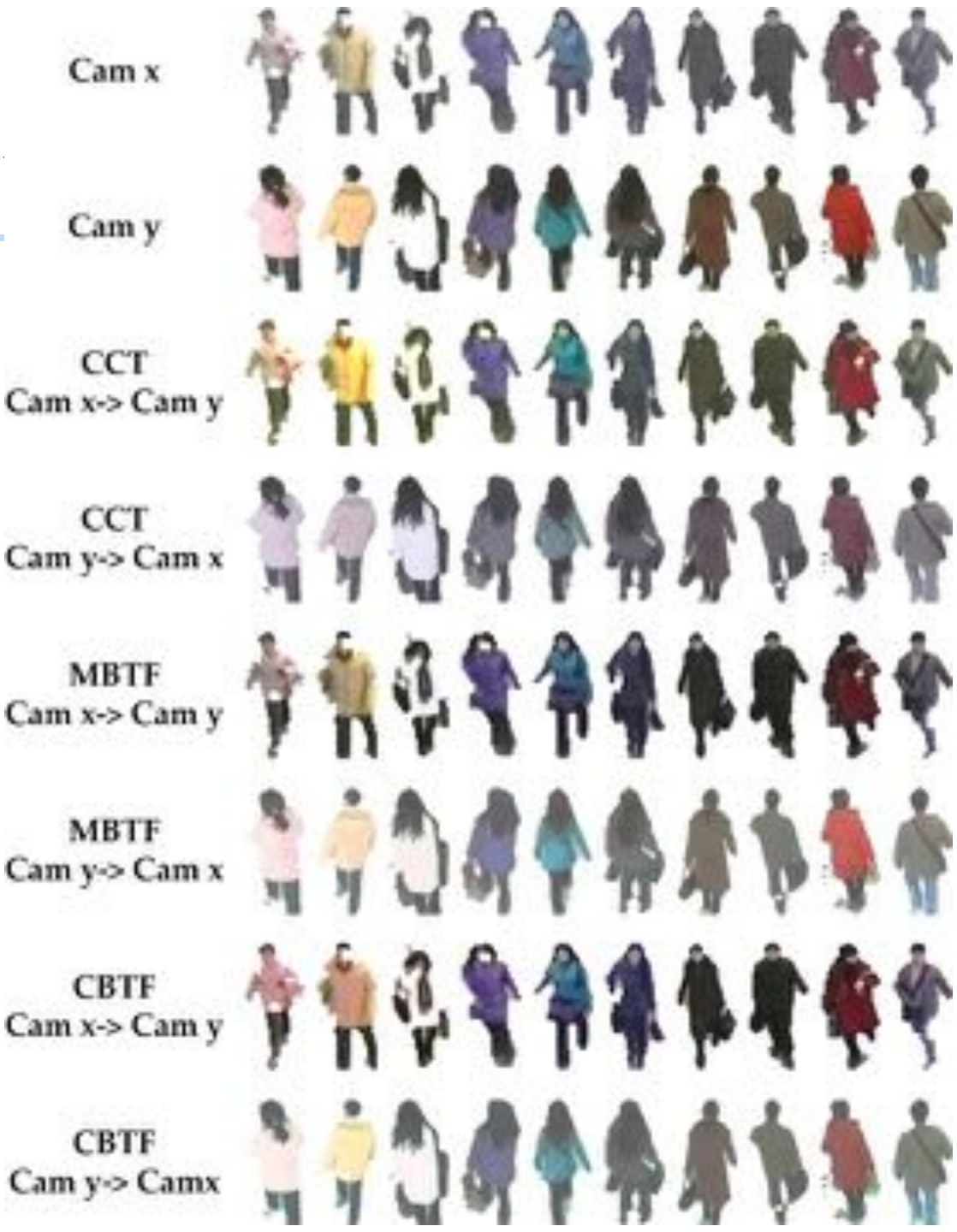
Under Cam_y

Experimental Results



(1) Performance is different with transfer directions.
(2) The proposed method (CCT) performs best as it does not depend on a large scale pre-labeled training dataset.

Some examples using different color transfer methods

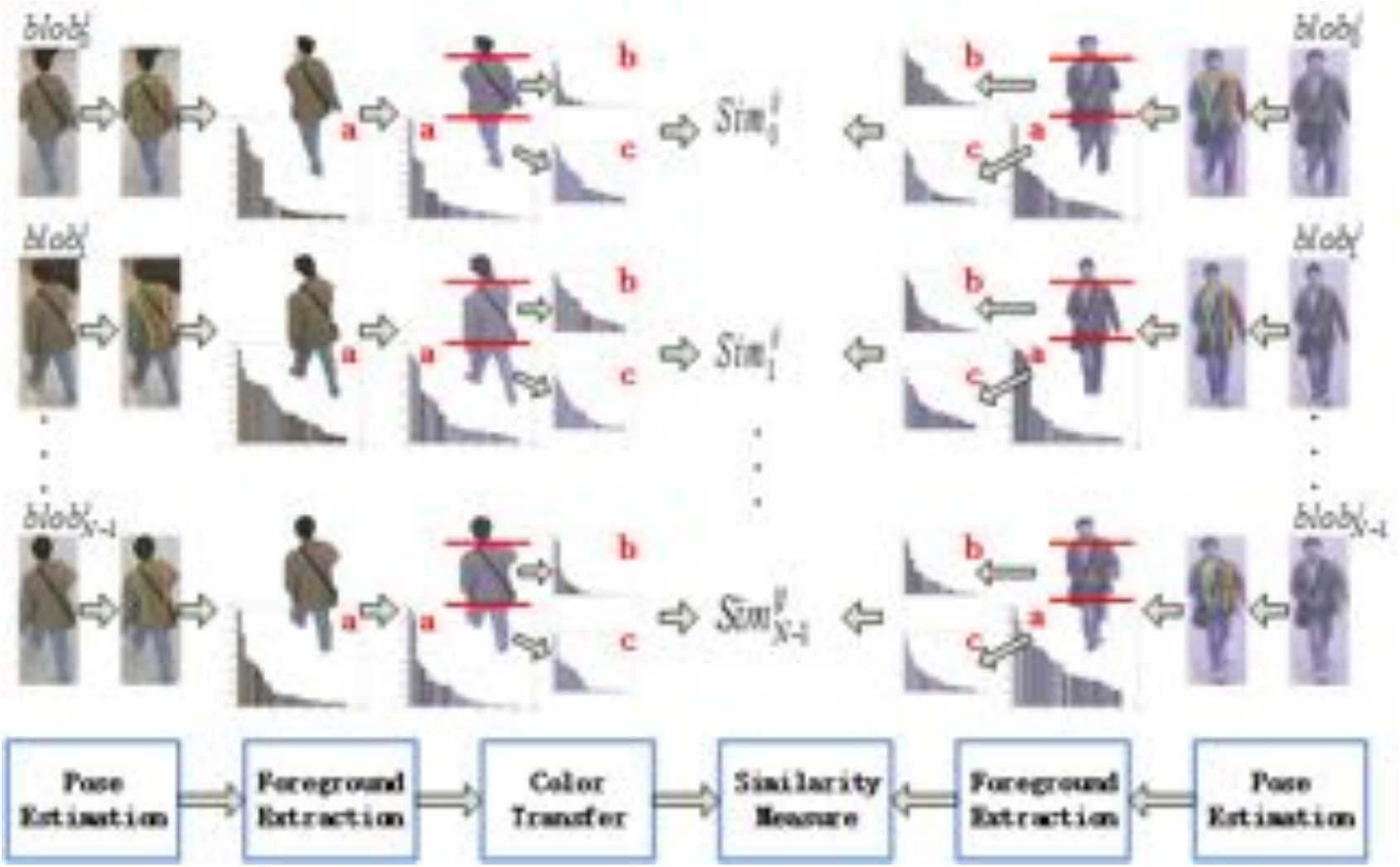


X.T. Chen, K.Q. Huang and T.N. Tan. Object tracking across non-overlapping views by learning inter-camera transfer models, PR, 2013.

3 Object Matching (Computing Object Similarity)

- **Problem**
  - **Object appearance varies considerably across cameras due to factors such as illumination, camera properties, viewpoints, poses and non-uniform clothing.**

Object Matching (Computing Object Similarity)

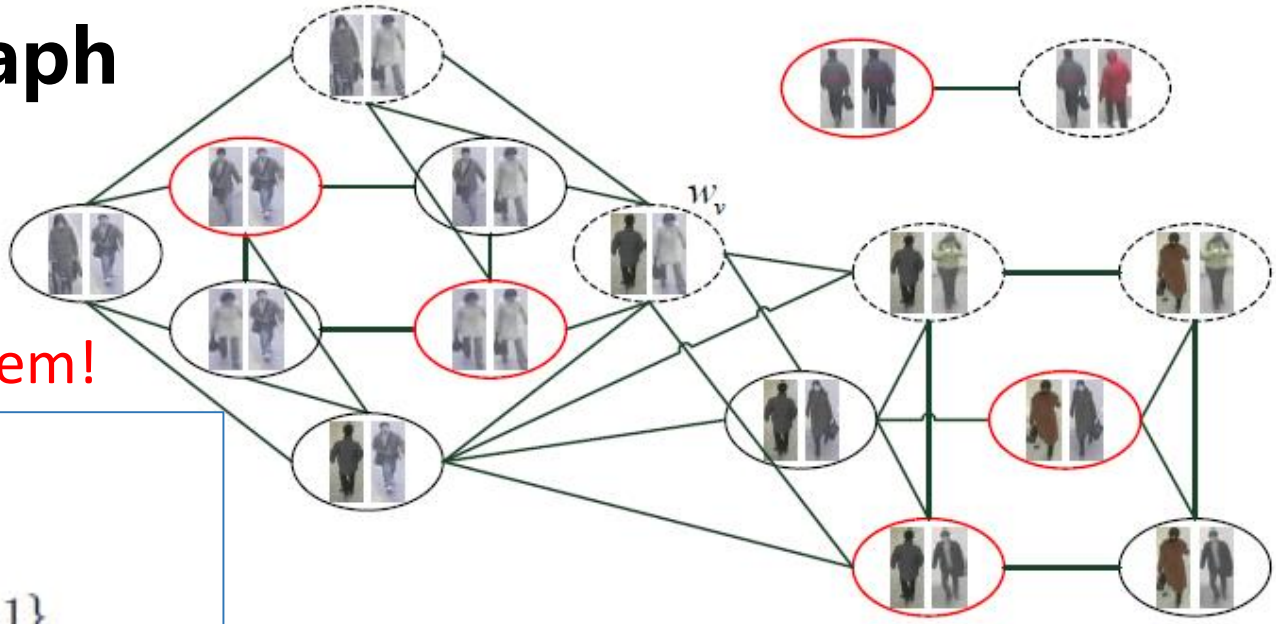


4 Find the MWIS Solution

- **Build the graph**
  - **Nodes:**
    - Candidate object pairs which are constrained by spatio-temporal cues (e.g. the connectivity and average transition time between two entry/exit zones)
  - **Edges:**
    - Two nodes are linked together when they contain a same object, no matter it is a departure object or an arrival object.
  - **Weights of Nodes:**
    - Similarities between the two objects in the same nodes

Find the MWIS Solution

- **Build the graph**



It is a NP-hard problem!

$$IP: \quad \max \sum_{v=1}^{N_V} w_v \chi_v$$
$$s.t. \ \forall v \in V, \quad \chi_v \in \{0,1\}$$
$$\forall (v,v') \in E, \quad \chi_v + \chi_{v'} \leq 1$$

N_v is the number of nodes in G_w;
w_v is the weight of Node v.

To find the MWIS solution:
Step1. find all the independent set
- an accelerated algorithm is presented to speed up the ergodic search of all the independent sets.

Step2. find the independent set with the maximum of the sum of weights

Experiments

- NLPR-MCT Dataset (http://mct.idealtest.org/)
  - Contain four datasets collected from different non-overlapping camera networks.
  - The ground truth of single camera object tracking results is given.
- Evaluation criteria
  - tracking accuracy of object tracking across cameras
  - $mcta = 1 - Mismatch_{cross}/TruePositive_{cross}$

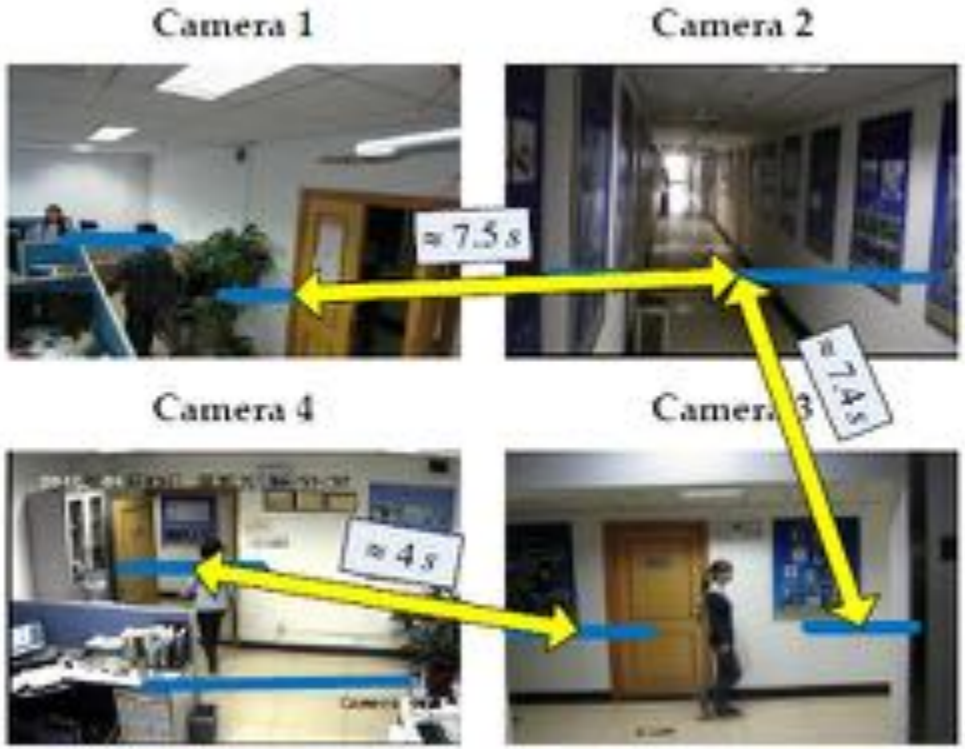
TRACKING PERFORMANCE ON THE NLPR-MCT DATASET.

mcta	Dataset 1	Dataset 2	Dataset 3	Dataset 4
MWIS-MCT [1]	0.925	0.875	0.855	0.762
Cai [2]	0.915	0.913	0.516	0.705

[1] X.T. Chen, J.G. Zhang, K.Q. Huang and T.N. Tan, Object Tracking across Non-overlapping Cameras by Finding Maximum Weight Independent Sets, submitted to TCSVT 2015.
[2] Y. Cai and G. Medioni, "Exploring context information for inter-camera multiple target tracking," in Proc. IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 761–768.

Tracking Results

- Four-camera network with non-overlapping views

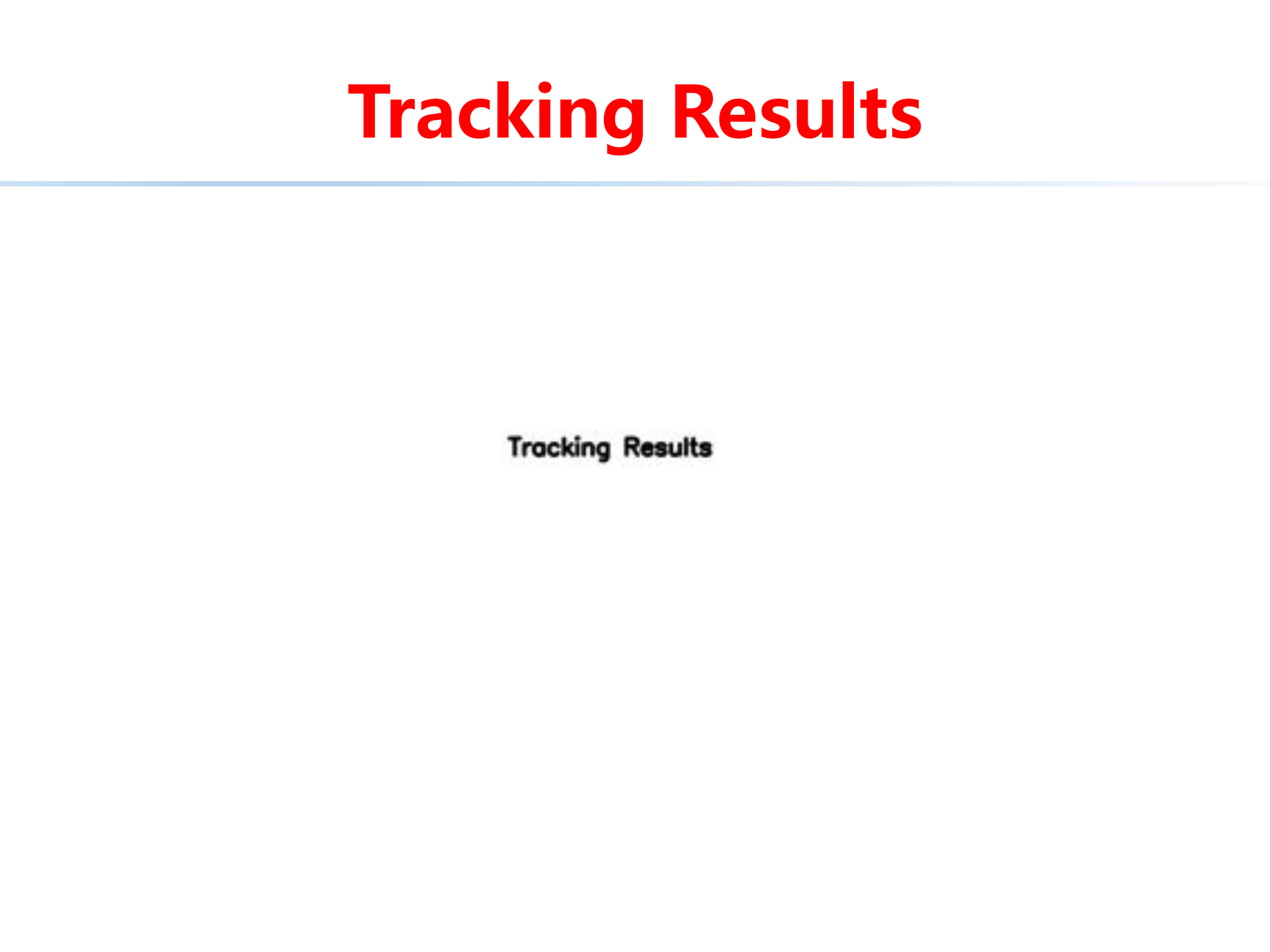


Tracking Results

Tracking Results

- Five-camera network with non-overlapping views

Tracking Results



Outline

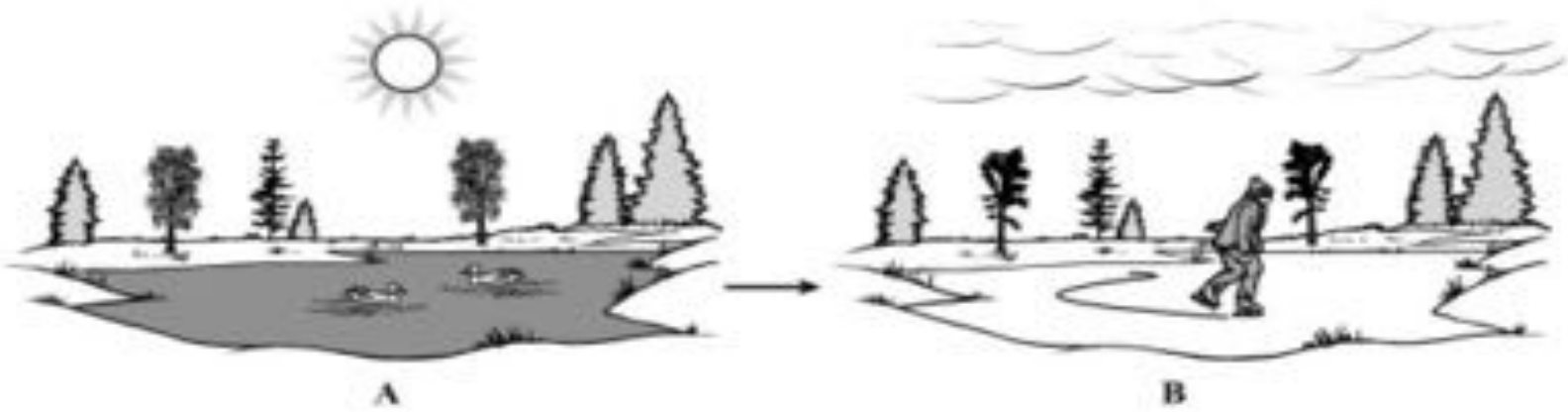
- Background and Context
- Feature Representation
- Model Learning
- LSVC: An Example
  - Large-Scale Visual Surveillance
- **Conclusions**

Conclusions

- **Exponential growth in visual data calls for large-scale visual computing (e.g. large-scale visual surveillance)**
- **Availability of big visual data facilitates deep learning, which in turn helps LSVC**
- **Despite of the significant progress made with deep learning, machine vision is still far away from human vision, especially for high-level visual tasks, e.g., semantic understanding of dynamic scenes.**

Conclusions

- What, Where, How, Why are the four key problems in computer vision. Currently, we have made great progress in **What and Where**, but not much on **How and Why**, which are even more important in large scale visual computing.
- Current research focuses on **"visible" appearance** of objects, but ignores **"logic" and "common sense"**, two very challenging and open problems for ultimate tasks of large scale visual computing.



Slide credit: Song-Chun Zhu

Thank You!