Computer Vision

Patch-based Object Recognition

Masataka Kagesawa

Contents

- Papers on Patch-based Object Recognition Using Images
  - This week and last week

- This week
  - Basic idea on recent object recognition
  - Comparison with 20Q
  - A paper presented in CVPR2007

What is “Object Recognition”?

- Traditional definition

  For an given object *A*, to determine automatically if *A* exists in an input image *X* and where *A* is located if *A* exists.

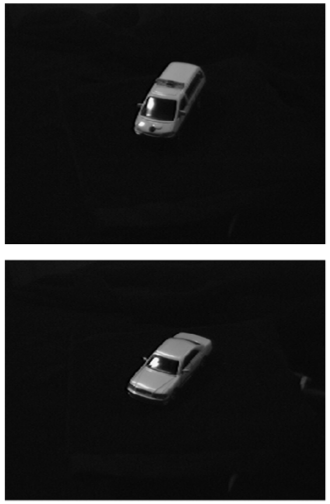
- Ultimate issue (unsolved)

  For an given input image *X*, to determine automatically what *X* is.

An example of traditional issue

- What is this car?
  - Is this car any of given cars in advance?

Input image



Training images

An example of ultimate issue

- What does this picture show?
  - Street, 4 lanes for each direction, divided road, keeping left, signalized intersection, daytime, in Tokyo,…

Recognition and Detection

- Recognition

  Example: biometric identification

  - Recognize you from your face image or so
- Detection

  Example: intruder detection

  - Detect objects whose temperature is around 37 degree C

- Recognition is much finer than detection

What is Recognition Target ?

- Specified an object
- Specified an object (unknown location, might be occluded)
- Any object of a specified class
  - You can define any class as you like
- Any object of any class

“Specified”: known features in advance

Recognize Specified Object(s)

- Give training images of the object(s)
- Make “model” (compressed database)
  - Robust against environment changes
- Search most similar model from an input image

Problem for traditional issue

Training image

Input image

Where is the left vehicle in the right picture?

How to make model

- Manual generation for each given object
  - Traditional
  - Camera-independent features

  →Environment-dependent features

  →Not very popular now

- Auto generation from training images
  - deductive method : PCA, SIFT as feature
  - inductive method : NN, GA

Requirement for model

- Independent from translation
- Independent from rotation
- Independent from scale
- Independent from environment

- Lower, more general but difficult

Structure of model

- Features from whole object are sensitive against environment
- Patch-based features are robust against environment
  - One patch-based feature is not enough
  - Model is defined as an configuration of lots of features.

20Q (break)

- Think of something and 20Q will read your mind by asking a few simple questions
- http://www.20q.net/

- This idea is the essence of recent patch-based object recognition

20Q as Object Recognition

- Targets: nouns (no proper nouns)
- Features/characteristic: yes-no questions

- Nouns are characterized as intersection of yes-no questions.
- 20 yes-no questions can recognize 2^{20} objects;
  - 2^{20} is about 1 million.
  - In OED, there are 0.3 million words
  - (World population: 10,000 million)

Discussion

- Fastest way: Sort words by dictionary order and ask with bisection method
  - Model of a word is its index number.
  - Index number is 1-dimentional.
- 20Q: each word is considered to belong in the intersection of the sets of given yes-no questions
  - Questions are manually created in advance
  - Model structure is "automatically" constructed

Interesting points in 20Q

- Answer to yes-no question may not be “yes” nor “no”.
- Some answers can be different from pre-learned answers.
  - Robust against environment
- Interactive
  - 20Q can select a question after it has the answer of the previous question.
  - 20Q can be supervised.

Difficulty on Object Recognition

- Give training images in advance
- Extract features from the images
  - Features: “yes-no” questions in 20Q
  - The questions must be automatically extracted
  - Answer is an operation result on the input image
  - Non-interactive: unsupervised

- What are good features?
- Answers might be probability.

Indoor and Outdoor

- Object recognition in outdoor is more complicated than that in indoor.
- Light
  - Indoor: controllable
  - Outdoor: uncontrollable
- Obstacles
  - Indoor: might be, can be removed
  - Outdoor: expected

Issues in Outdoor

Basic Technique (1)　(review)

- An Image is considered as a vector.
- BW image of 256x256, 8bit depth can be one of $(256*256)^{256}=2^{4096} \fallingdotseq 10^{1300}$
- Using whole image is not practical
  - One digital camera image can be mega-pixel; $((1M)^{256})^3$ =? (about 10^{4500})
- Model should be compact

Basic Technique (2)

- Still image or image sequence (movie) ?
  - Movie: rich information
  - Still image: finer image
  - Method which work on still images can work on image sequences

  - Trade-off: movies are popular now.

Basic Technique (3)

- Is camera fixed or moving?
  - Fixed: Is camera location and pose known?
    Yes, usually can be calibrated
  - Moving: Is camera motion known?
    No, usually but yes sometimes.

- Does environment of target objects change?
  - Do target objects move? (fixed location, rotation, scale?)
  - Is light source controllable? (fixed shade, fixed shadow?)

Basic Technique (4)

- Database from training images
  - Smaller, better (# of all qs must be small)
  - Larger, longer matching time (20Q→30Q)
- Supervised method?
  - Non-supervised method is better

Basic Technique (5)

- There might be several answers in the end
  - Still going on: they are just candidates
- Hierarchical method
  - First question in 20Q; not yes-no question
  - Narrow down candidates and find optimal one.

Recent Technique (1)

- Probability and lots of Questions
  - Bag-of-Features

    Q: how many this feature are there in the object?

    A: number or probability

    Distribution of the answers becomes the model
  - Ada boost

    Each question is foolish; sure to divide two

    Understand the characteristic of each question

    Lots of questions (>>20) identify the object
- Number is power!

Recent Technique (2)

- Big data
  - How to treat?
- Point cloud
  - Organized or not?
- Deep Learning
  - What is theory?

How to deal with “big data”

- No definitive theory yet
- Two research types:
  - No theory but somehow it works good
  - Nice theory but few examples
- Here, take theoretical approach

Paper review (1)

- PEET: Prototype Embedding and Embedding Transition for Matching Vehicles over Disparate Viewpoints
- *Yanlin Guo Ying Shan Harpreet Sawhney Rakesh Kumar*
- Sarnoff Corporation (USA)
- CVPR 2007

Objective

Figure 1. Top Row: A single object viewed by different cameras in disparate locations exhibits large appearance change. Middle & Bottom Rows: A single object viewed by multiple cameras in disparate locations and various orientations exhibits large pose change.

- Propose PEET, which can identify the same vehicles viewed by different cameras shown in the left figures.

Assumptions

- Take image sequences on fixed cameras
- Each vehicle can be tracked in each sequence
- The types of vehicles are given as 3D CG

(undocumented assumptions)

- Camera position and pose against road is known
- Cars run in almost constant speed
- Car scale is fixed (no lane changes)

Overview of PEET

- PE(Prototype Embedding)
  - Find the most similar *N1* models from One track sequence from Camera 1
- ET(Embedding Transition)
  - For each model, convert track sequence from Camera 2
- Model-to-image: select candidates
  - Select similar *N2* image sequences viewed by Camera 2
- Final answer
  - Optimal match among *N1**N2* combinations

Overview

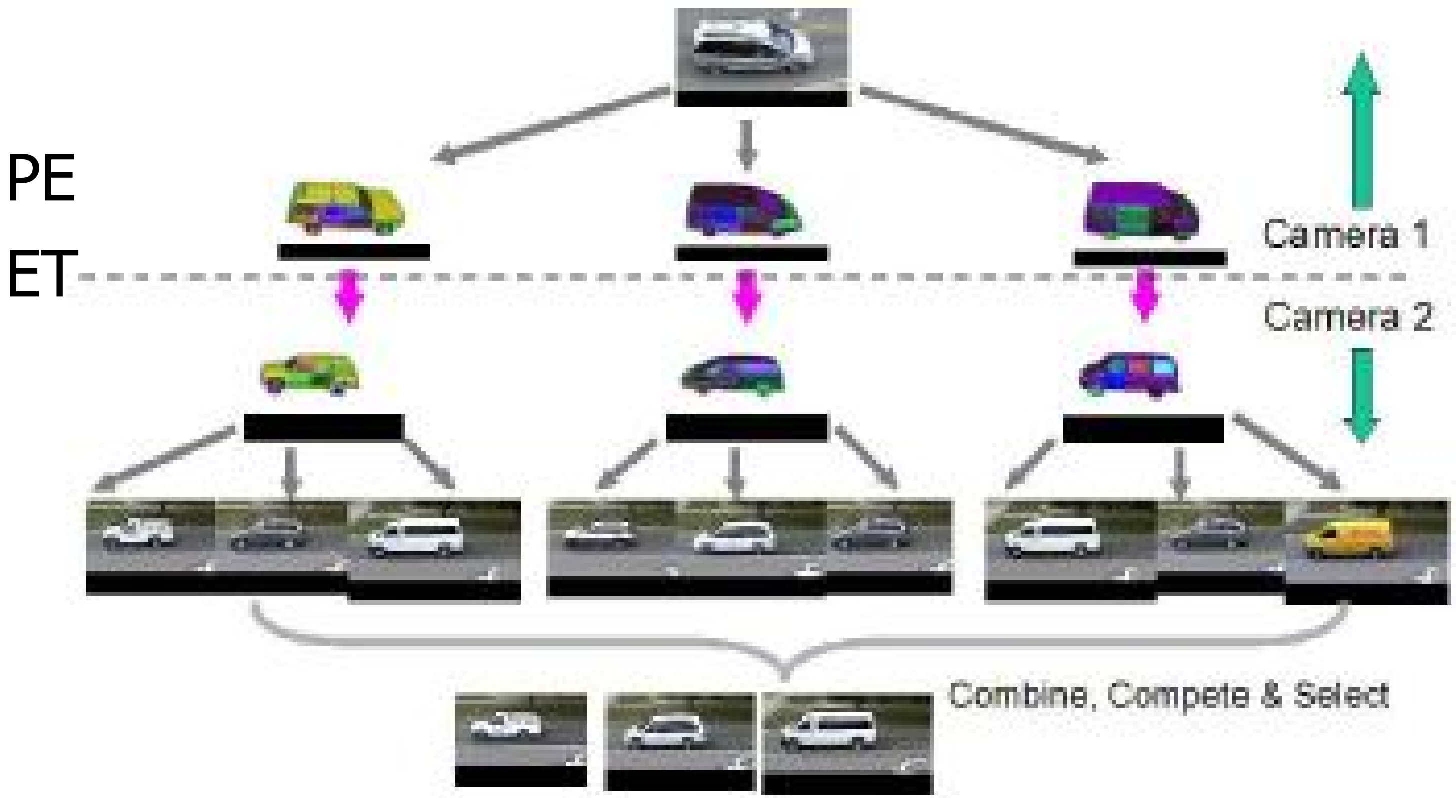


Figure 2. Overall schema of PEET.

Model

- K –dimentional vector, each component is the difference of k-th frame and the first frame

$d_{i,j,k}$: difference between k-th frame of Object i viewed by camera j and original image

For each i,j, $(d_{i,j,1},\ldots,d_{i,j,k})$ is the model of track sequence of object i viewed by camera j

Specification of this model

- Compare with image size, *K* is small.
  - One second, 30fps, then *K*=30-dimentional
  - Vehicle area: even 10x10, 100-dimentional
- Use edge image instead of original
  - Do not consider the difference of colors

- Model to vehicle is not 1-to-1.
- Models of similar vehicles are similar

Similarity of model

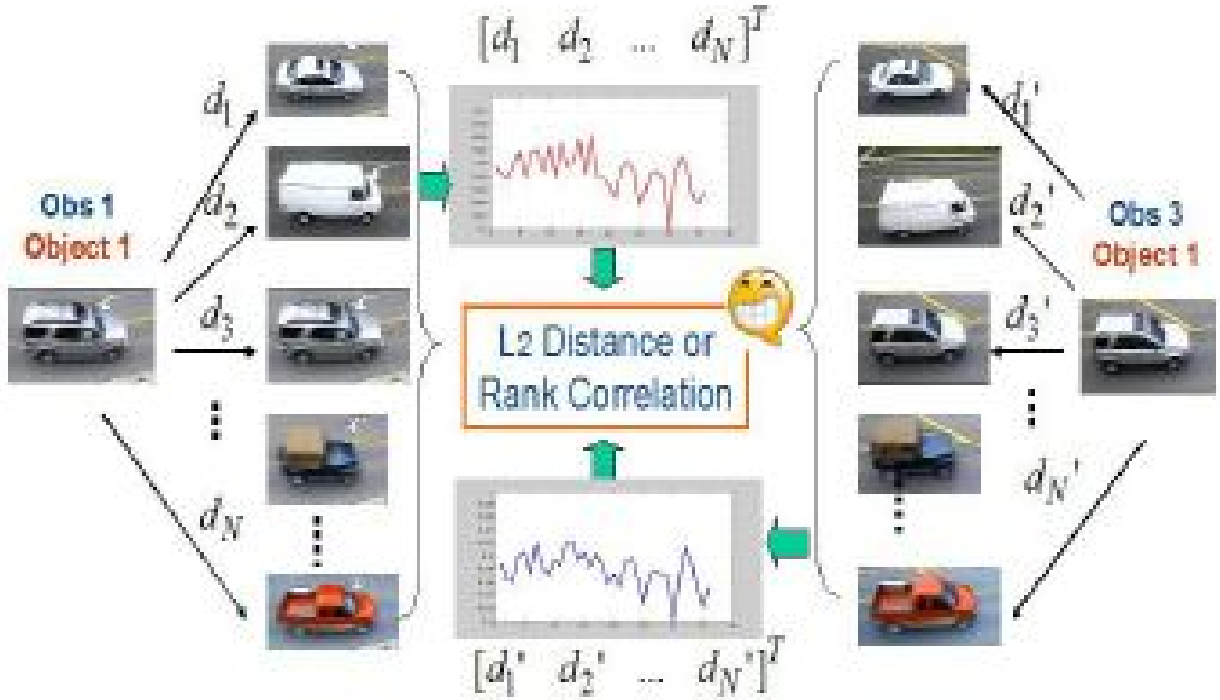


Figure 3. Image exemplar based embedding illustration. (For simplicity, subscripts denoting object and view indices are omitted in the distance representation.)

Recognition with this model

- Assume that views by camera 1 and camera 2 is similar
- K Questions:

  For each object i viewed by camera 1 and object j viewed by camera 2,

  Is $d_{i,1,1}$ and $d_{j,2,1}$ is similar?

  Is $d_{i,1,2}$ and $d_{j,2,2}$ is similar?

  …

  Is $d_{i,1,K}$ and $d_{j,2,K}$ is similar?

Problem on this method

- Need a lot of comparison (d x d')
- Sensitive against different environment of two cameras

- No good for different car pose.
  - If camera 1 views car front and camera 2 views car rear, then no similarity among models in camera 1 and models in camera 2

Failure Example

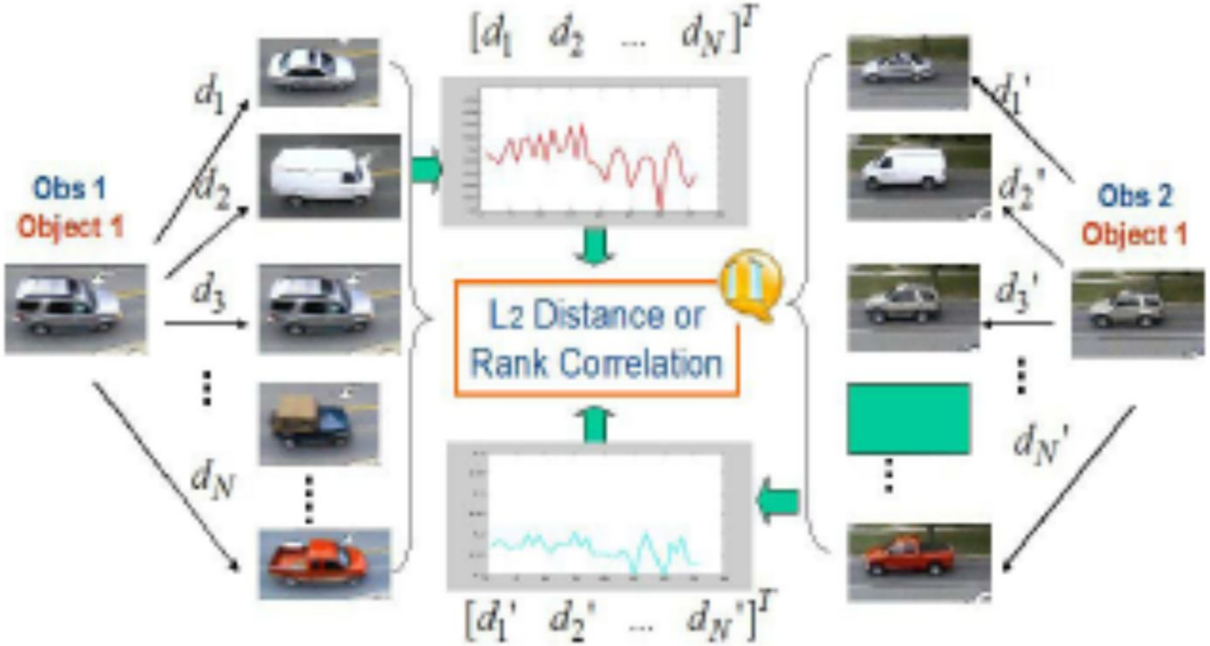


Figure 4. Exemplar Embedding cannot match objects with large pose change in this example. A complex mapping function needs to be computed between the embedding distances from the two cameras.

PE(Prototype Embedding)

- Prepare 3D CG models of vehicles
- Each CG is colored so that it is easy to extract edges

- External camera parameter is known
- For each CG i and camera j, $d_{i,j}$ is calculated in advance.
- We call $\{d_{i,j}\}$'s PE.

Edge Extraction from CG

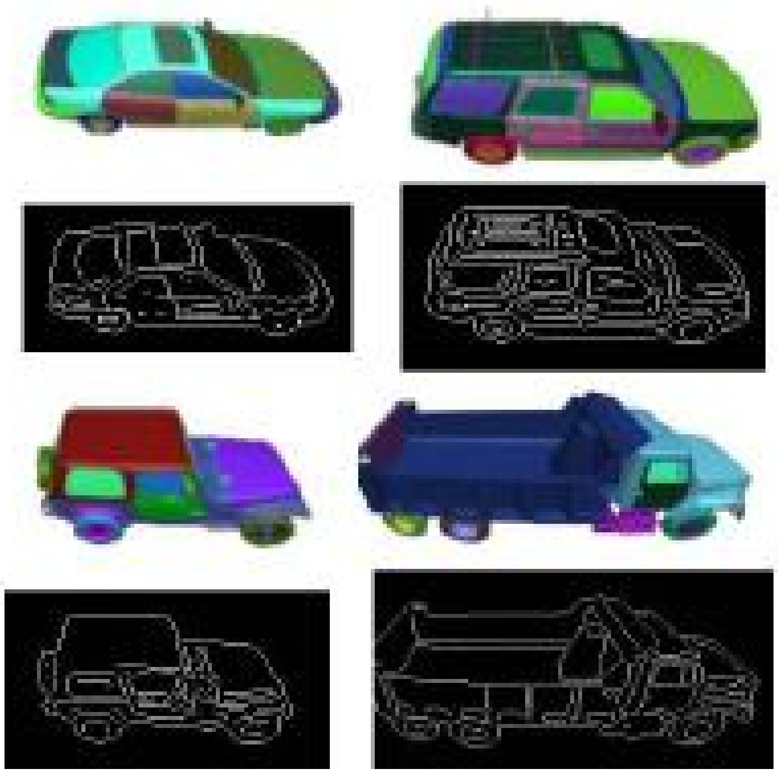


Figure 5. Some representative vehicle prototypes and their edge maps.

ET(Embedding Transition)

- External camera parameters are known
- Image sequence of camera 1 → $d1,I$ (PE)
- $d2,I$ (PE) → Image sequence of camera 2

- Using PE, we can compare $d1,j$ with $d2,j'$

Similarity of PE

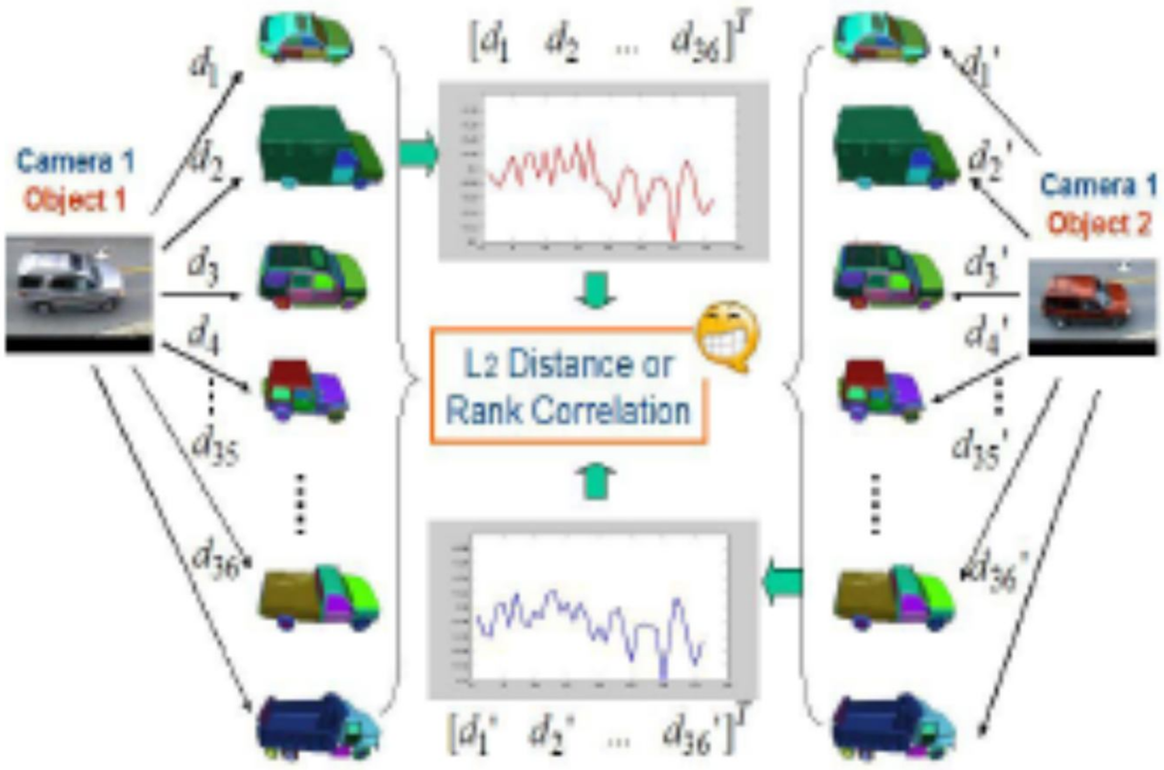


Figure 6. A Schematic of Prototype Embedding.

Vehicle Class Recognition on PE

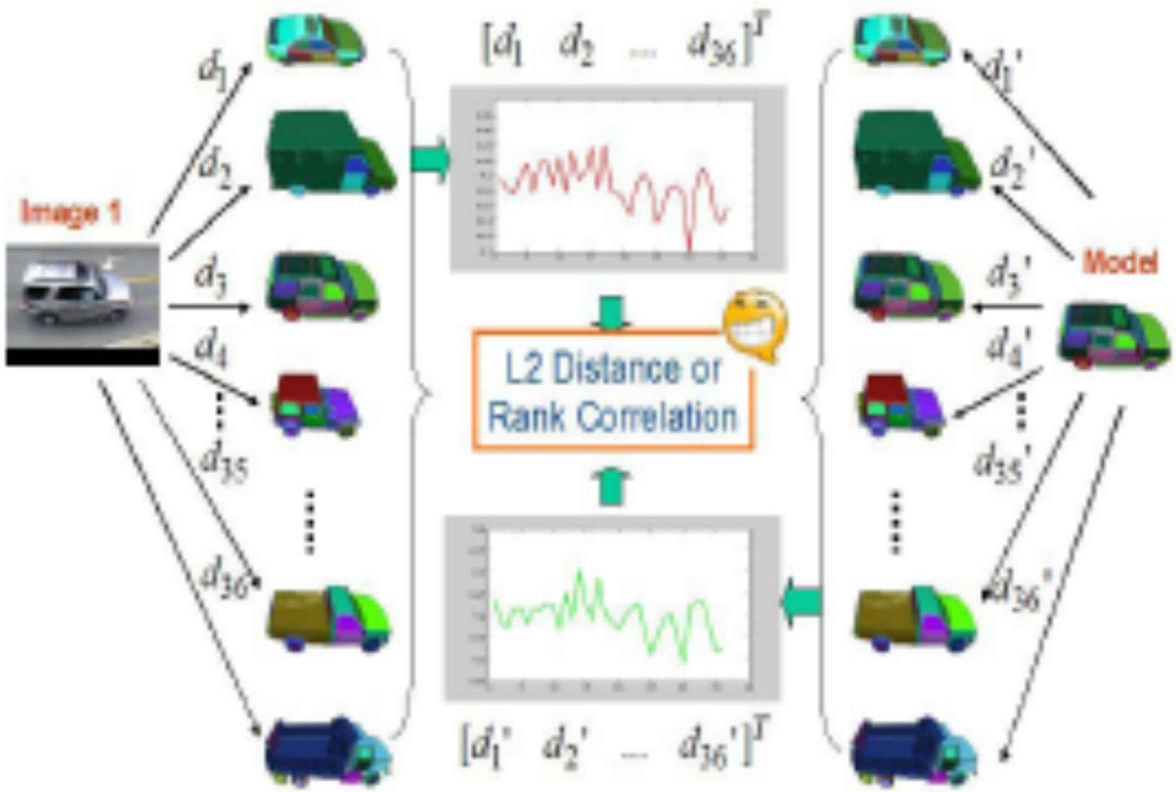


Figure 7. A Schematic of Model embedding.

Justification of PE

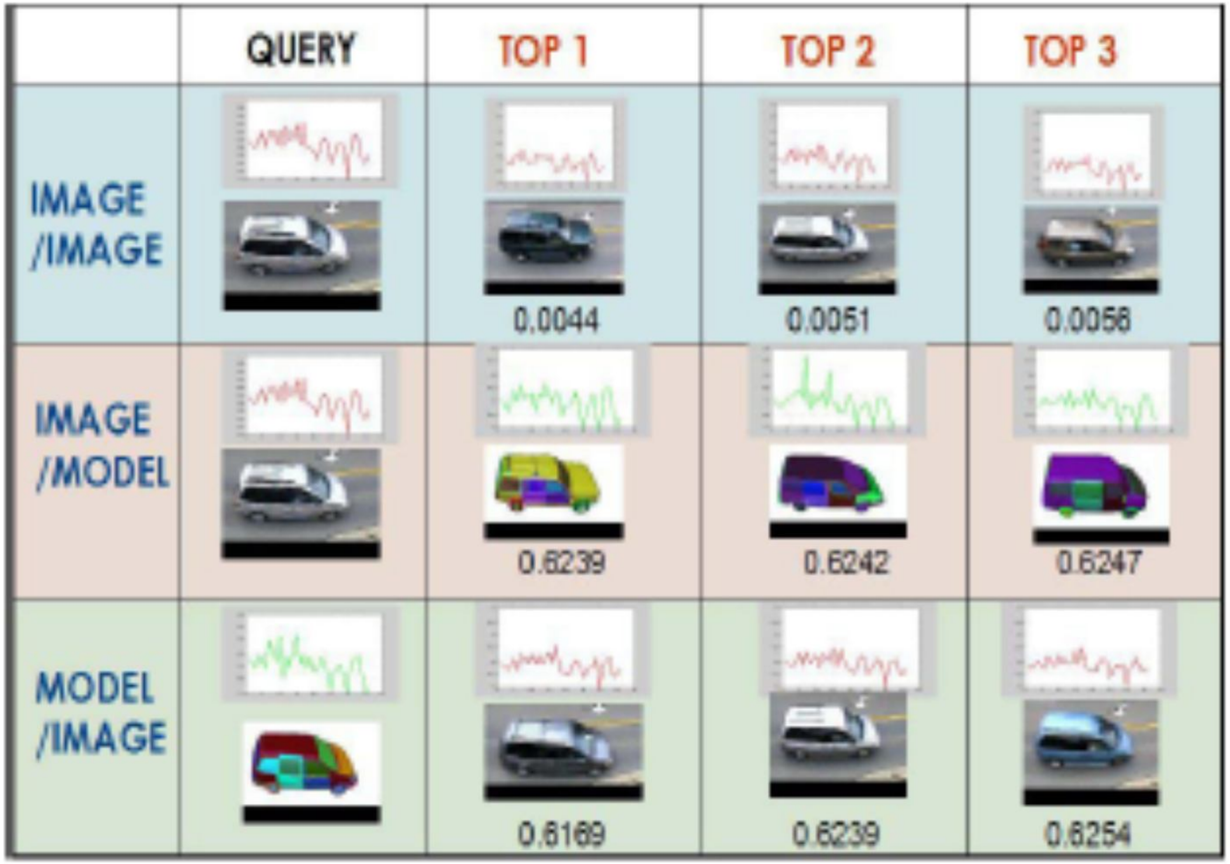


Figure 8. Model-Image embedding transition example.

Improvement with symmetry

- PEET so far
  - camera 1 image →camera 1 CG model (PE)

    →camera 2 CG model (ET)

    match camera 2 image
  - One-way
- PEET new
  - candidates→camera 1 CG model (ET again)

    match camera 1 image

    Select matches original sequence only

New PEET works anytime?

- It works fine if the resolution of two cameras is almost the same (or the size of bounding box of target objects are almost the same)
- It does not work if the resolutions of two cameras are different
  - What to do?
  - Use RBF.

Different Resolution Case

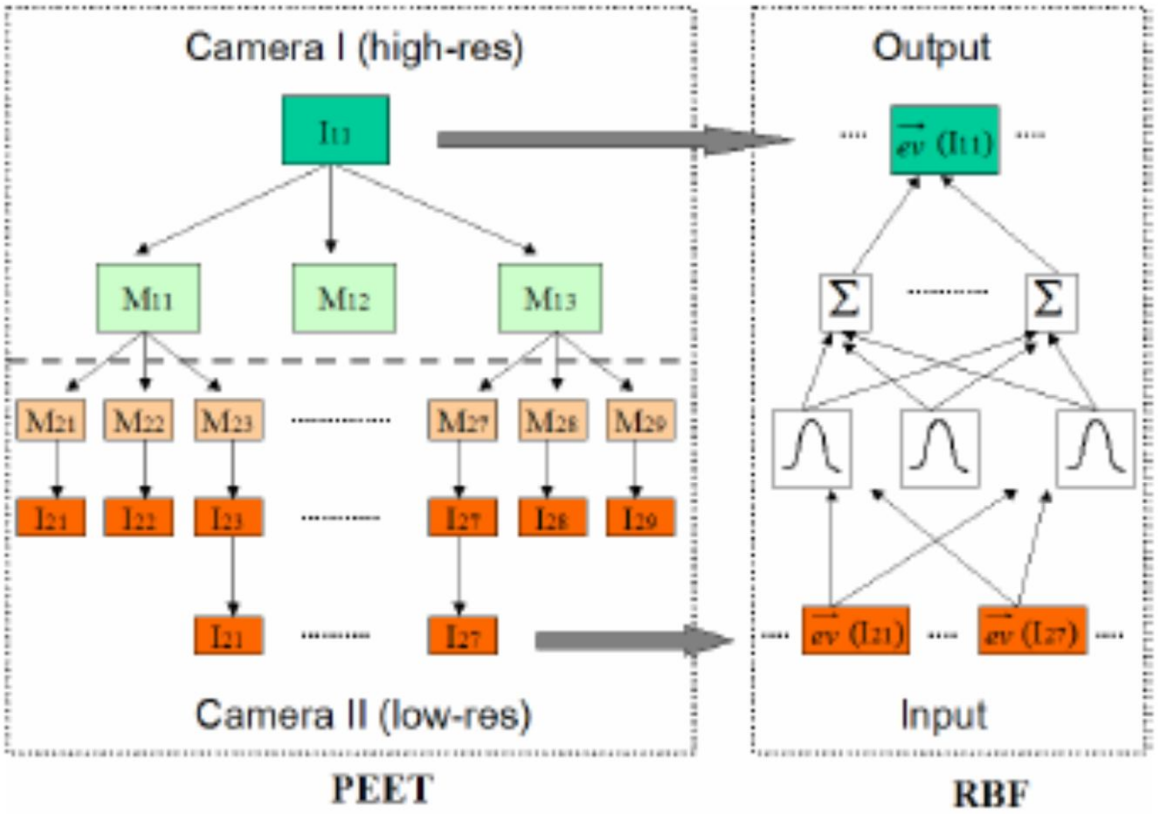


Figure 9. Un-supervised Learning with PEET.

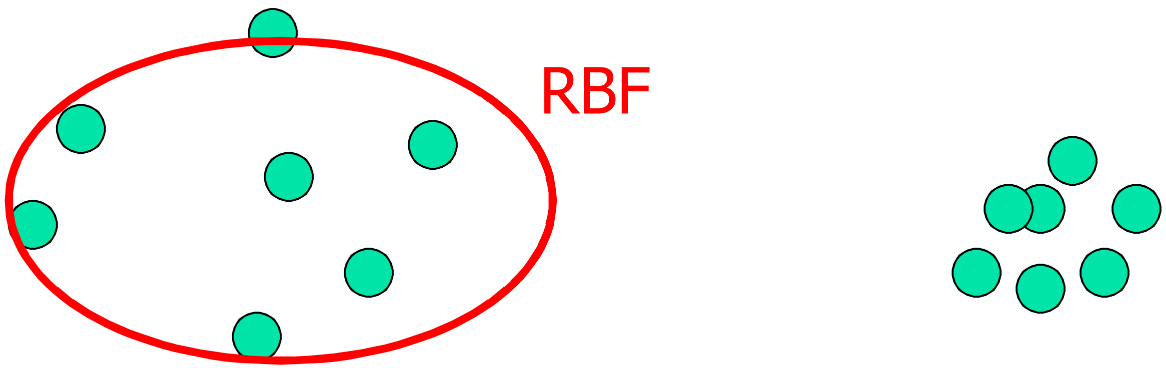
Explanation

- Camera 1: high resolution
- Camera 2: low resolution
- Camera 2 model is considered as a "deformation" of camera 1 model
- RBF: is a function which shows degree of deformation

- RBF (Radical Basis Function): is obtained from camera 2 CG models.

Rough explanation

K-dimentional space

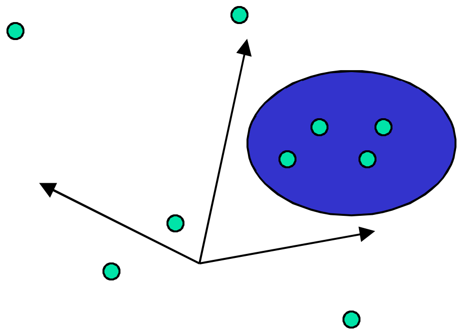


Low resolution

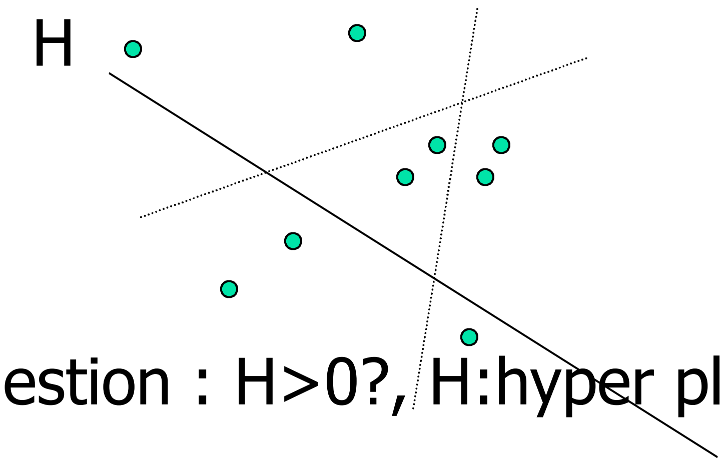
High resolution

Class Recognition

- RBF



- 20Q (SVM)



one question : H>0?, H:hyper plane

Points of PEET

- Vehicle CGs are prepared in advance
- Feature is a point in *K*-dim vector space
  - One object track to vector
  - One image to one number
  - *K*-questions will distinguish the target.
- Match two sequences in different poses
  - This kind of task is usually very hard

Similarity in two cameras (ET)

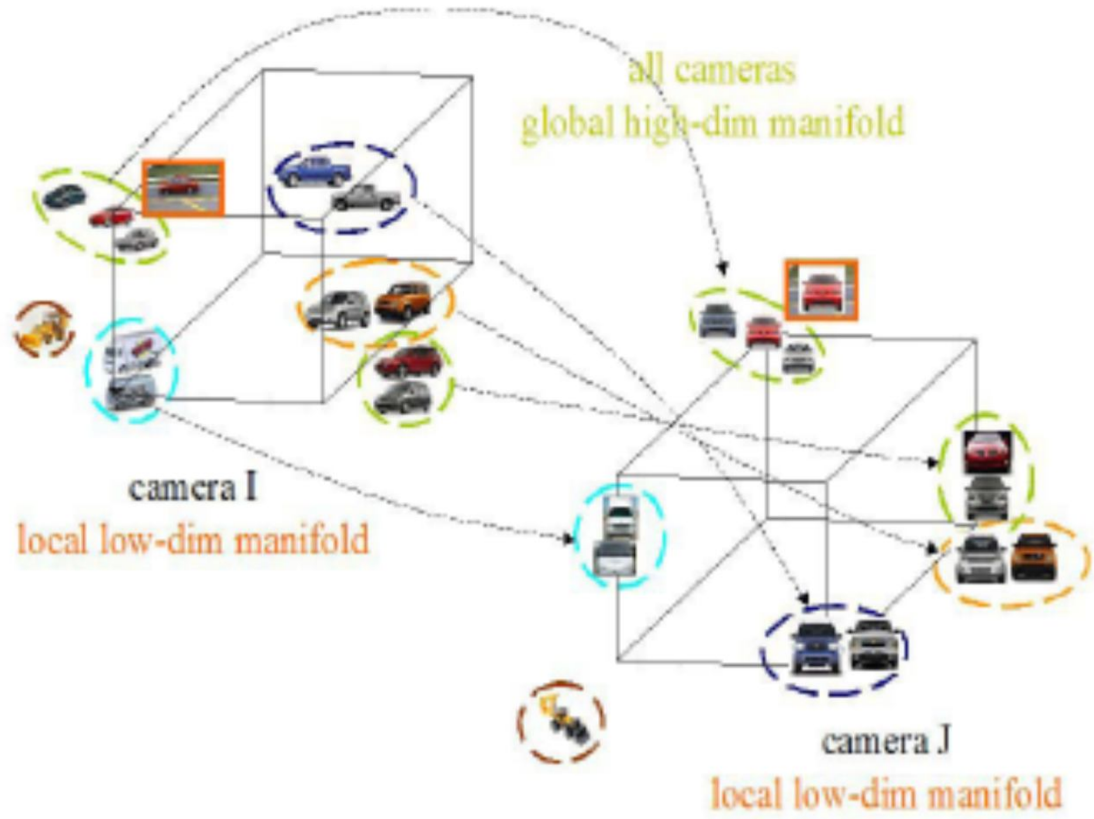


Figure 10. Space tessellation using prototype models.

Correspondence of 2 cameras

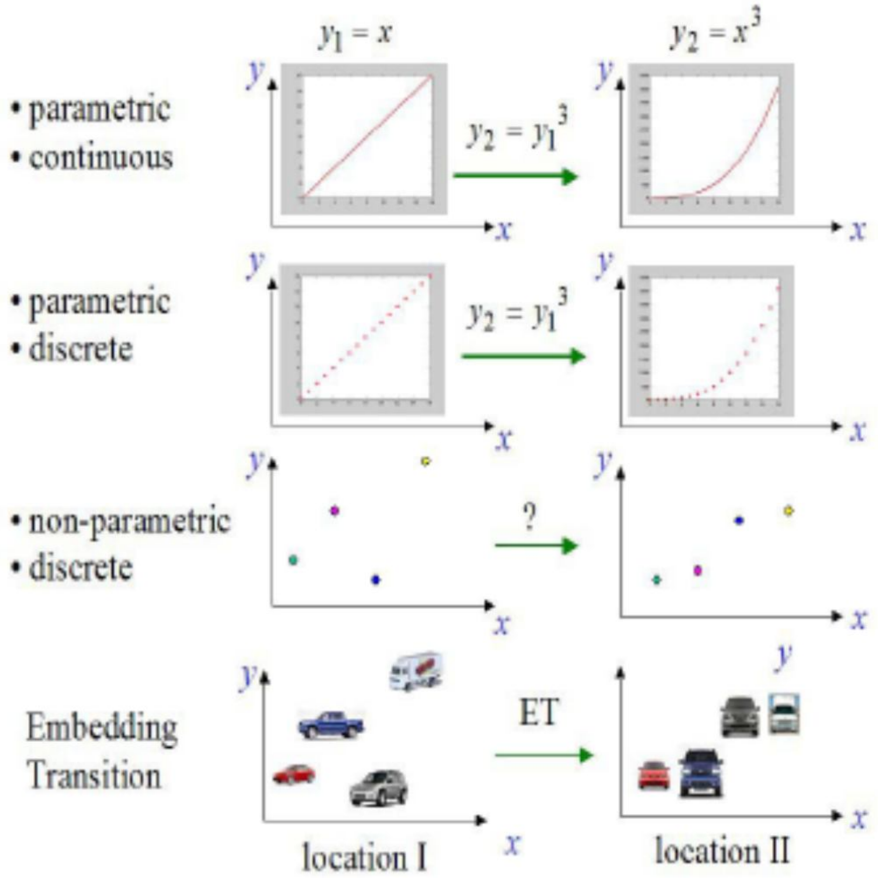


Figure 11. Embedding Transition (ET) as non-parametric discrete function mapping.

Applications of PEET

- Class recognition using PE
  - Case of high resolution camera
  - Case of low resolution camera

- Matching between images on different cameras whose location and pose are different

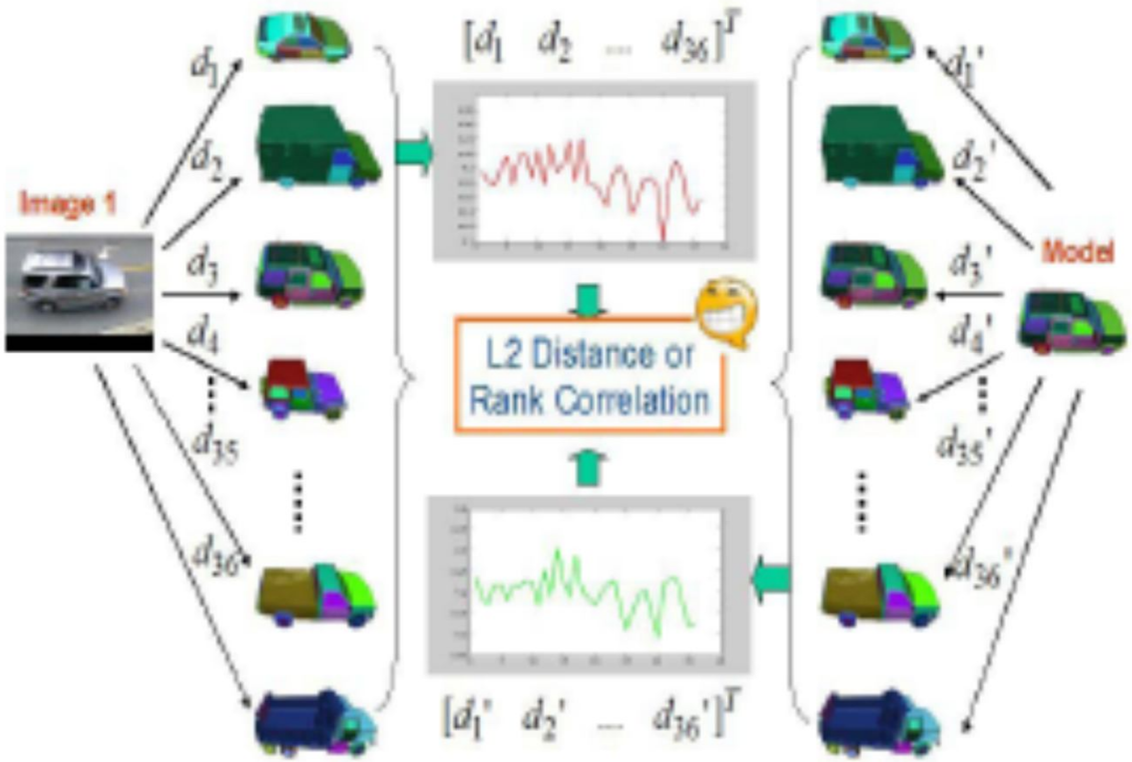
Experiments

- Traffic monitoring cameras spread in area of $4km^2$
- Each road has 2-3 lanes/direction.
- Video image of 30min. Length (traffic volume is 200 vehicles/30min)

- High-res: close lane from camera
- Low-res: far lane from camera (0.5-0.9)

Class recognition on PE(hi-res)

- Image→model



Class recognition on PE(hi-res)

- Data set 1

TD=(Si)/(detected Si)
MD=(missed Si)/(total vehicles)

Table 1. True & Overall Detection Rates for DS1

	TD(S1)	TD (S2)	TD (S3)	TD (S4)	TD_all (S)
cam1	87.65%	96.55%	62.50%	93.33%	88.82%
cam3	86.59%	94.92%	47.37%	100.00%	85.96%
cam7	82.80%	89.06%	72.73%	100.00%	85.39%
cam11	92.59%	85.51%	83.33%	100.00%	89.56%

Table 2. Miss Detection Rate for DS1

	MD (S1)	MD(S2)	MD (S3)	MD(S4)
cam1	0%	13.85%	9%	39.13%
cam3	2.74%	13.85%	0.00%	54.17%
cam7	0.00%	18.57%	33.30%	50.00%
cam11	1.32%	7.81%	25.00%	36.00%

S1:Sedan
S2:mini van
S3:one box
S4:pick up

# of S3, S4 is small

Class recognition on PE(hi-res)

- Data set 2

TD=(Si)/(detected Si)
MD=(missed Si)/(total vehicles)

Table 3. True & Overall Detection Rates for DS2

	TD(S1)	TD (S2)	TD (S3)	TD (S4)	TD all (S)
cam1	97.52%	86.09%	46.67%	87.80%	90.06%
cam3	94.37%	86.21%	42.86%	96.55%	89.34%
cam7	97.35%	83.33%	63.64%	96.30%	90.12%
cam11	95.19%	76.53%	50.00%	85.71%	84.62%
cam15	94.23%	84.21%	58.33%	95.24%	88.73%

Table 4. Miss Detection Rate for DS2

	MD (S1)	MD(S2)	MD (S3)	MD(S4)
cam1	7.65%	10.81%	30.00%	12.20%
cam3	2.90%	15.73%	0.00%	28.21%
cam7	5.98%	6.59%	0	31.58%
cam11	16.10%	15.73%	0.00%	14.29%
cam15	2.00%	13.51%	0.00%	37.50%

S1:Sedan
S2:mini van
S3:one box
S4:pick up

# of S3, S4 is small

Class recognition on PE(lo-res)

- Image→model + RBF

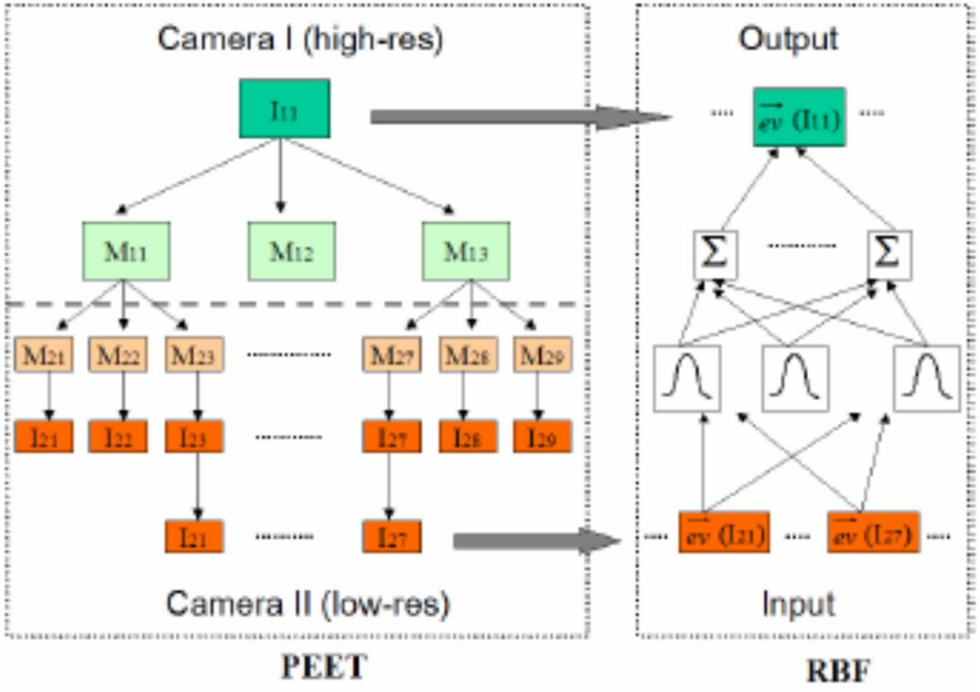


Figure 9. Un-supervised Learning with PEET.

Class recognition on PE(lo-res)

- Result

TD=(Si)/(detected Si)
MD=(missed Si)/(total vehicles)

Table 5. Far Lane Object Classification Performance Comparison w/ & w/o Learning Based Mapping

	Cam 2		Cam 4	
	NEW	OLD	NEW	OLD
TD (S1)	90.54%	60.56%	88.75%	91.43%
MD (S1)	6.94%	18.87%	2.74%	33.33%
TD (S2)	80.00%	57.14%	90.00%	63.64%
MD (S2)	15.79%	48.94%	16.67%	6.67%
TD (S3)	100.00%	100%	70.00%	75.00%
MD (S4)	50.09%	86.96%	25.00%	52.63%
TD (S4)	80.95%	58.33%	100.00%	87.50%
MD (S4)	26.09%	46.15%	25.00%	22.22%

S1:Sedan
S2:mini van
S3:one box
S4:pick up

# of S3, S4 is small

Matching between two cameras

- Image→model→image & v.v.

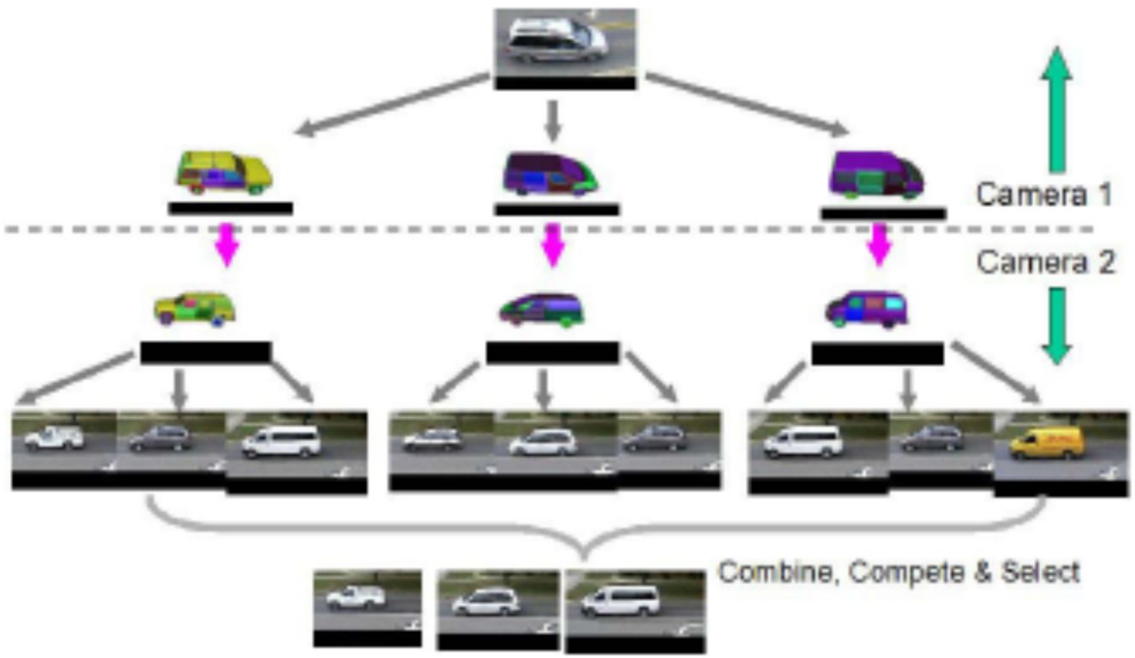


Figure 2. Overall schema of PEET.

Result (1)

Figure12. Demonstration of object querying. The leftmost column shows the vehicle images used as queries. Each of the corresponding rows on the right show the vehicle objects returned as matches ordered from best to worst.

Result (2)

Figure12. Demonstration of object querying. The leftmost column shows the vehicle images used as queries. Each of the corresponding rows on the right show the vehicle objects returned as matches ordered from best to worst.

Matching result

Table 6. Object Query Performance for Both Same and Different Side Objects

Cross Camera Query for Same Side Lanes		Cross Camera Query for Different Side Lanes	
	Accuracy		Accuracy
cam001-003	97.63%	cam001-002	93.60%
cam001-007	97.25%	cam008-011	88.00%
cam011-015	97.87%	cam003-016	94.44%
cam004-002	95.18%	cam004-007	91.02%
cam012-008	95.79%	cam001-012	94.06%

Technical point in this paper

- Model from outdoor image sequence
  - Edge-based image
- Image sequence processing
  - One image to one number
- Correspondence in different resolution
  - RBF is adopted
- Correspondence in different poses
  - CG (ET) is proposed

Comparison with 20Q

- Edge-based outdoor image
  - Accuracy of the answer gets good
- One image to one number
  - Automatic generation of questions
- RBF is adopted
  - Theoretical background for fuzzy answer
- CG (ET) is proposed
  - Consistency of different questions

Vehicle Identification Method

- Other vehicle identification methods are proposed matching vehicle sequences
- This method does not seem to be good for vehicle identification
- License plate reading system, vehicle-to-roadside communication system are in practical in Japan

Summary

- Essence of object recognition
- Using 20Q…
  - An configuration of lots of feature is unique
  - How to generate "good" features
  - How robust the features are
  - Answer can be probability
- Theoretical approach on "big data"

Preview

- **Semantic Hierarchies for Recognizing Objects and Parts**
  - Boris Epshtein Shimon Ullman
  - *Weizmann Institute of Science, ISRAEL*
- **Accurate Object Localization with Shape Masks**
  - Marcin Marszaek Cordelia Schmid
  - *INRIA, LEAR - LJK*