

Актуальные задачи и методы современного компьютерного зрения (анализ материалов конференций PCV'14 и ECCV'14)

Визильтер Ю.В., viz@gosnias.ru



European Conference on Computer Vision

Photogrammetric Computer Vision – PCV 2014

ISPRS Technical Commission III Midterm Symposium
5th – 7th September 2014, Zurich, Switzerland

Конференция ECCV'14

ECCV 2014 registration and budget

- 1462 people registered online
(incl. 100+ workshop only)
- 46 people registered on site until yesterday
- About 100 student/postdoc helpers and chairs

- Total budget about 1MCHF
- About 100kCHF corporate sponsoring

Конференция ECCV'14

ECCV 2014 - Some Numbers

- 1444 complete submissions
 - ▶ thanks to all authors !
- 363 (27%) papers accepted
 - ▶ 325 (24%) accepted as posters
 - ▶ 38 (3%) accepted for oral presentation
 - ▶ (85 rejected without review)
- Multi-stage process
 - ▶ over 1000 reviewers
 - ▶ 53 Area Chairs

Основные темы, ключевые слова и тенденции

- Convolution networks, Deep learning, Sparse coding, Image Retrieval
- Structure-from-Motion, SLAM, 3DFlow
- 3D-data, RGBD-data, 4D-data
- Segmentation, Video Segmentation, MRF, CRF, Energy, Graphical Models, Superpixels
- Multi-Tracking, Human/Pedestrian Detection
- Human pose estimation, Human Action Detection and Prediction
- Crowd behavior, Group analysis
- Saliency
- Part-Based Deformable Models
- Face Detection, Alignment, Recognition
- Shape Analysis, Laplace-Beltrami operators, Manifolds
- Symmetric Positive Definite (SPD) matrices, Riemannian Manifolds, Grassmann manifolds
- Benchmarks
- UAVs, Fish-eye cameras, Unmanned Cars

Отдельные интересные работы и результаты

- Picture and Video Annotation, Image Tags, Video Tags
- Deblurring
- Image Matching and Fusion
- Super-resolution
- Intelligent Lighting

Содержание презентации

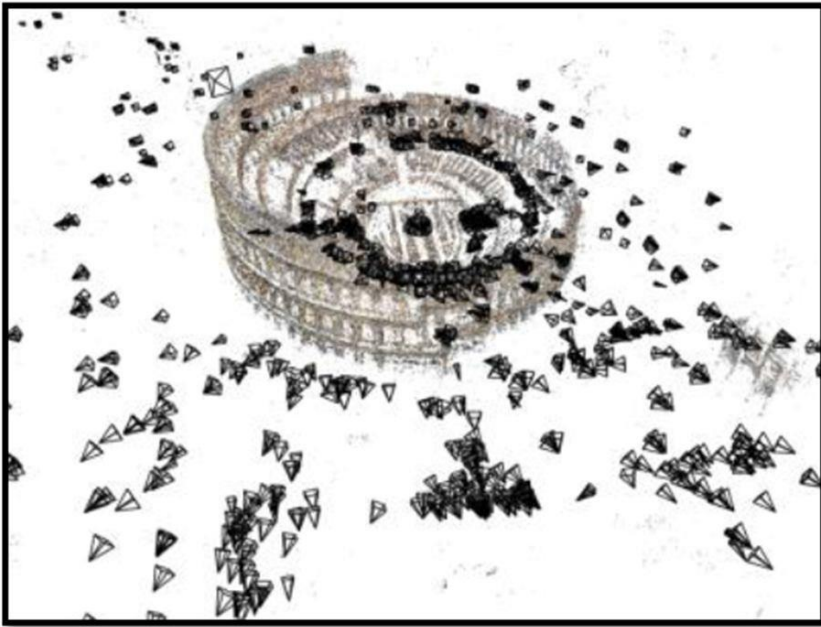
- **Реконструкция 3D сцены и навигация в ней:** Structure-from-Motion, SLAM, Road Scene Understanding and Autonomous Driving
- **Понимание типового видеосюжета:** Multi-Tracking, Human Detection, Human pose estimation, Human Action Detection and Prediction, Crowd behavior, Group analysis, Face Detection and Recognition
- **Распознавание изображений:** Convolution networks, Deep learning, Sparse coding, Image Retrieval
- **Зрительное внимание:** Saliency
- **Сегментация:** Image, Video & 3D Segmentation, MRF, Energy-based, Graphical Models, Superpixels, 3D-Flow
- **Форма:** Morphology, Shape Analysis, Manifolds, SPD
- **Контроль результатов и уровень задач:** Benchmarks

Реконструкция 3D сцены и навигация в ней:

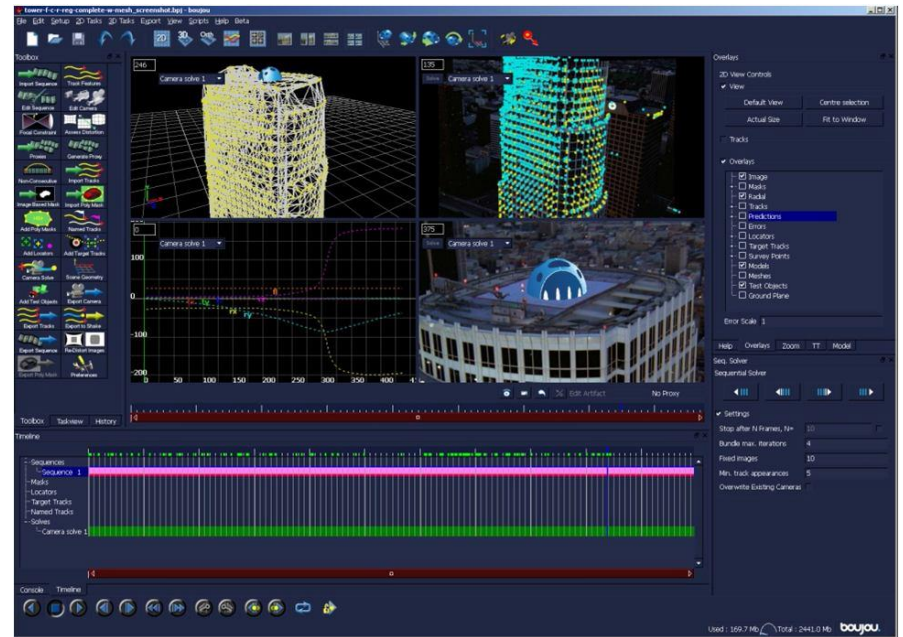
Structure-from-Motion,
SLAM, Road Scene
Understanding and
Autonomous Driving

Structure-from-Motion

- Structure-from-Motion – технология реконструкции 3D сцены на основе множества разноракурсных снимков и оценки положения/параметров относительной ориентации снимков



Building Rome in a day.

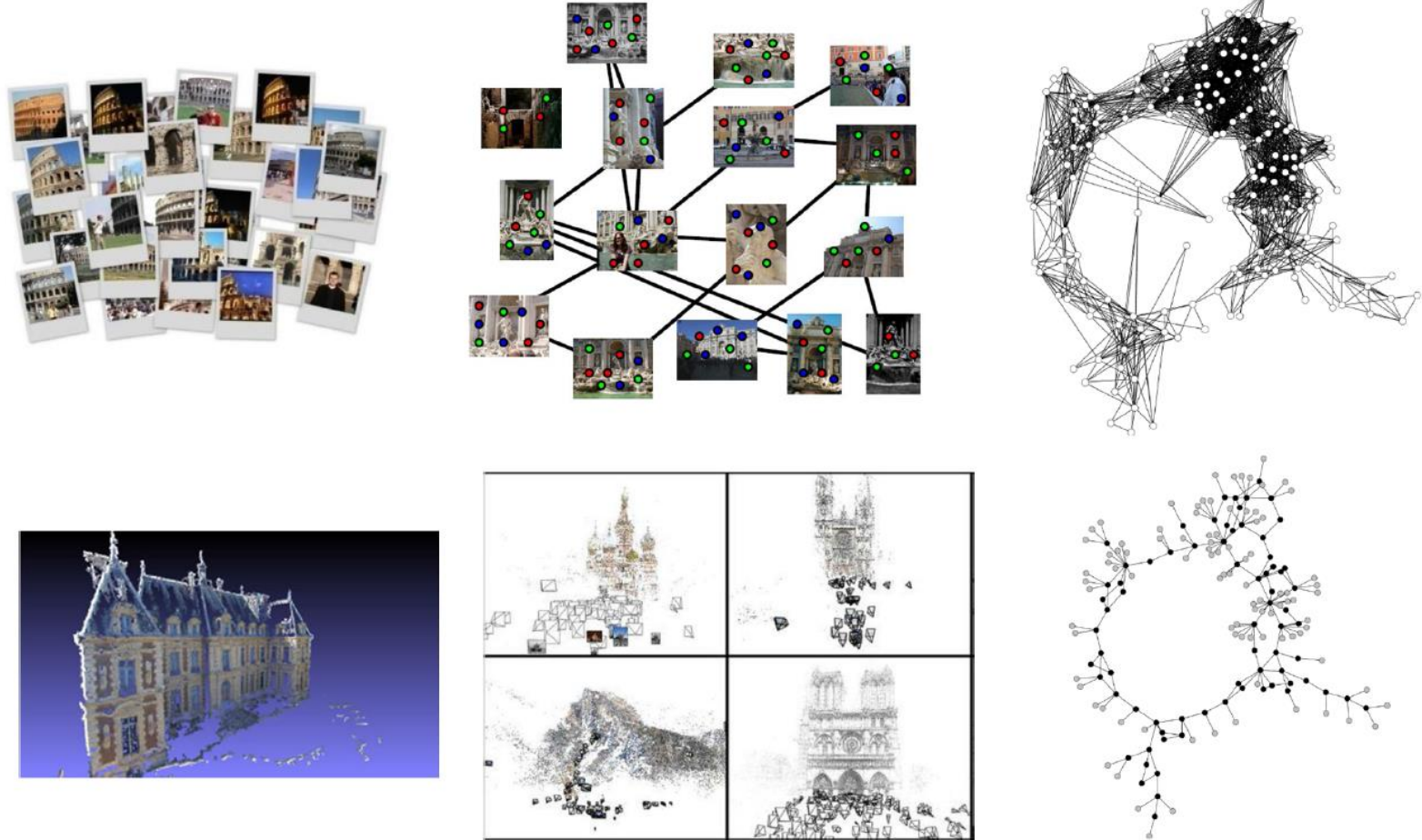


ICCV 2009 Boujou

- Библиотеки – Bundler, CMVS, PMVS и т.д.

Structure-from-Motion

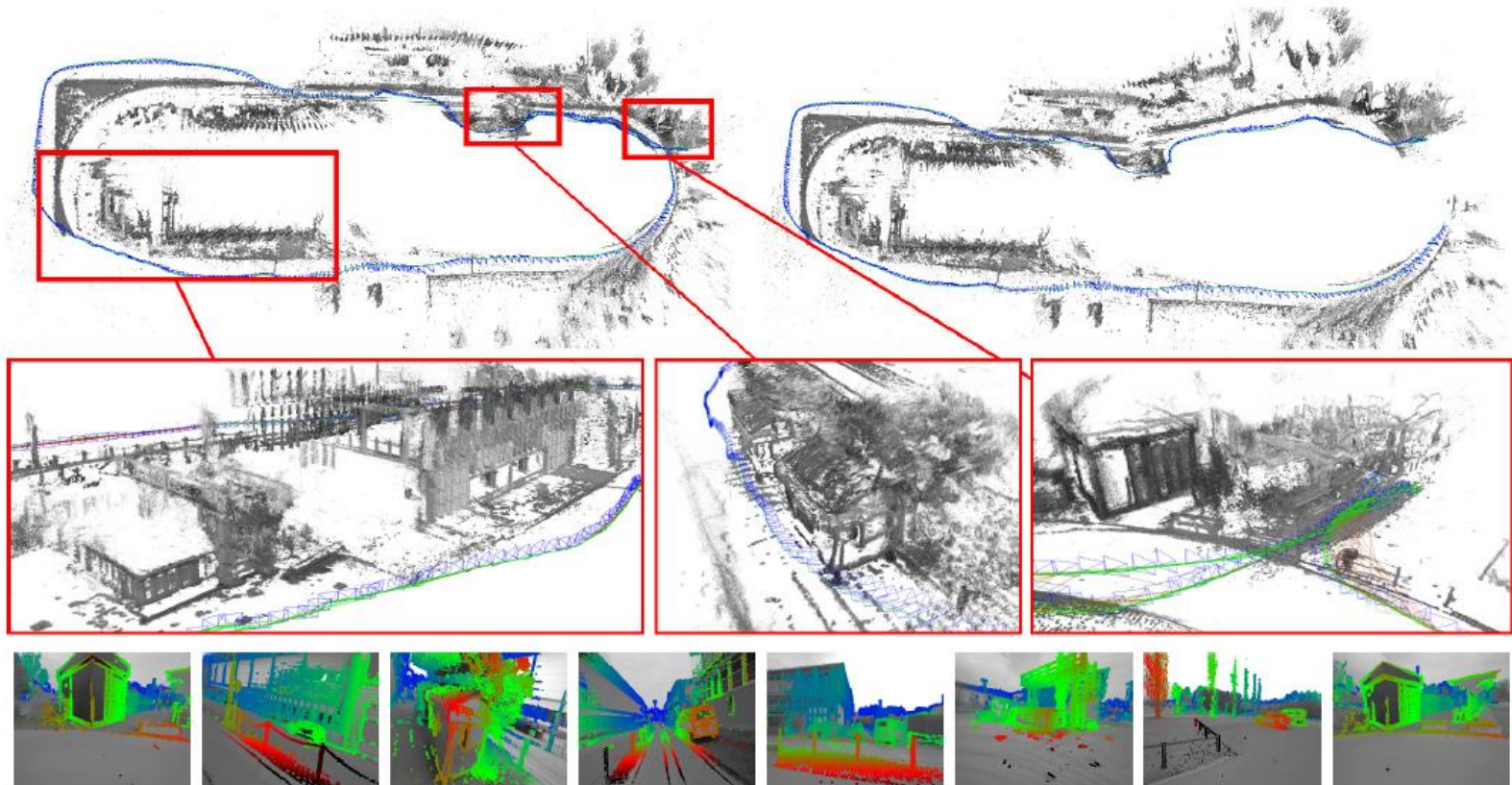
- Structure-from-Motion – технология реконструкции 3D сцены на основе множества разноракурсных снимков и оценки положения/параметров относительной ориентации снимков



Scene Chronology. Kevin Matzen, Noah Snavely. ECCV2014

SLAM

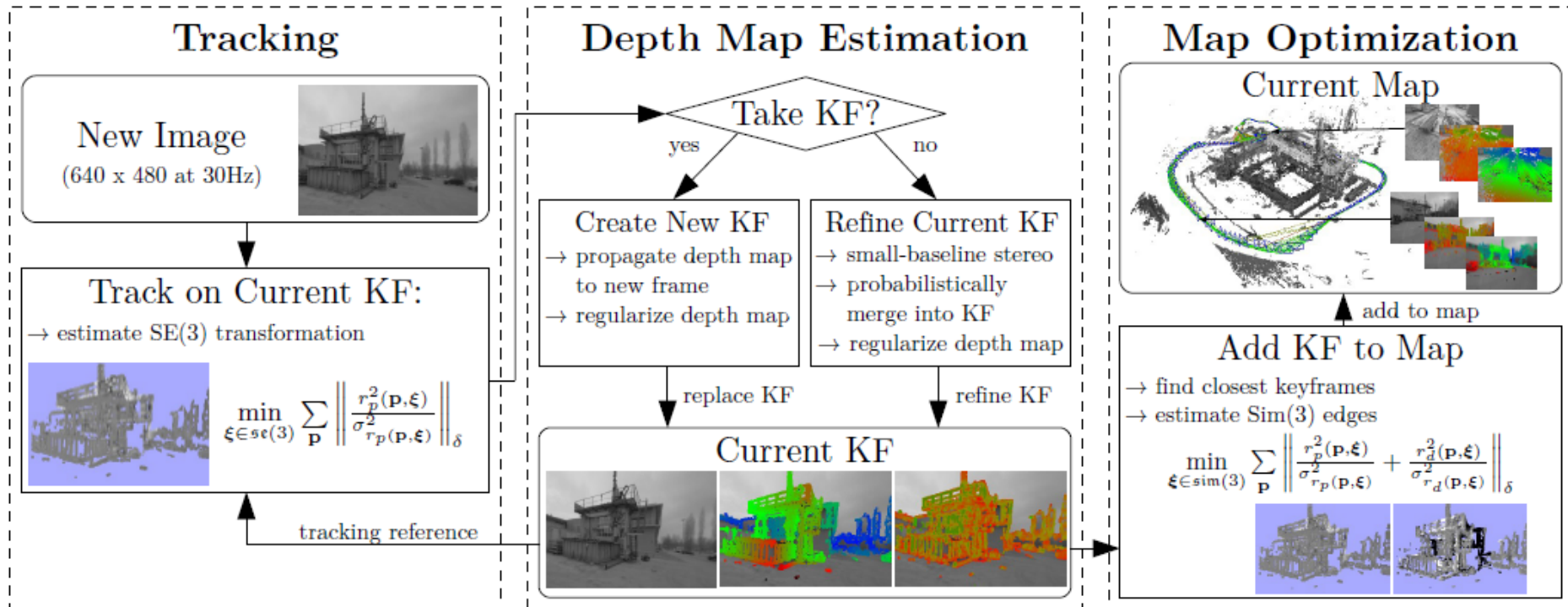
- SLAM – технология одновременной реконструкции 3D сцены и оценки положения/параметров движения камеры



LSD-SLAM: Large-Scale Direct Monocular SLAM,
Jakob Engel and Thomas Schops and Daniel Cremers, ECCV'14

SLAM

- SLAM – технология одновременной реконструкции 3D сцены и оценки положения/параметров движения камеры

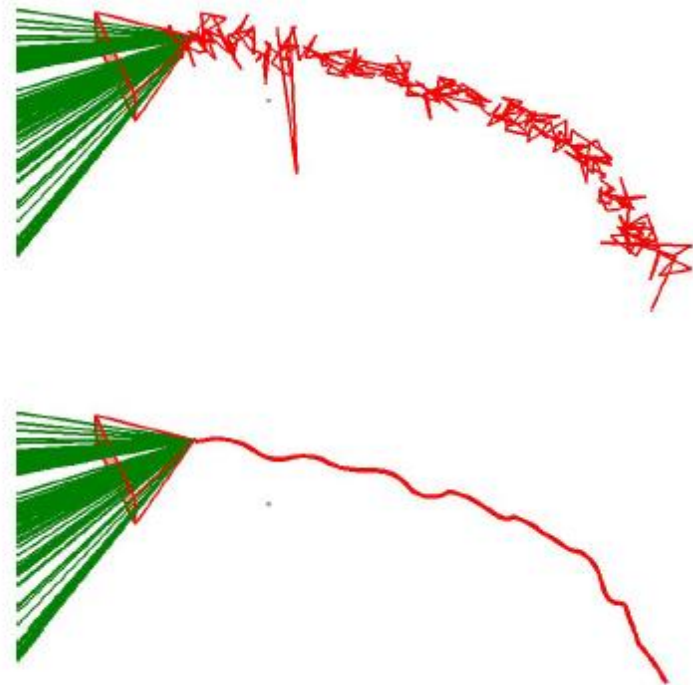
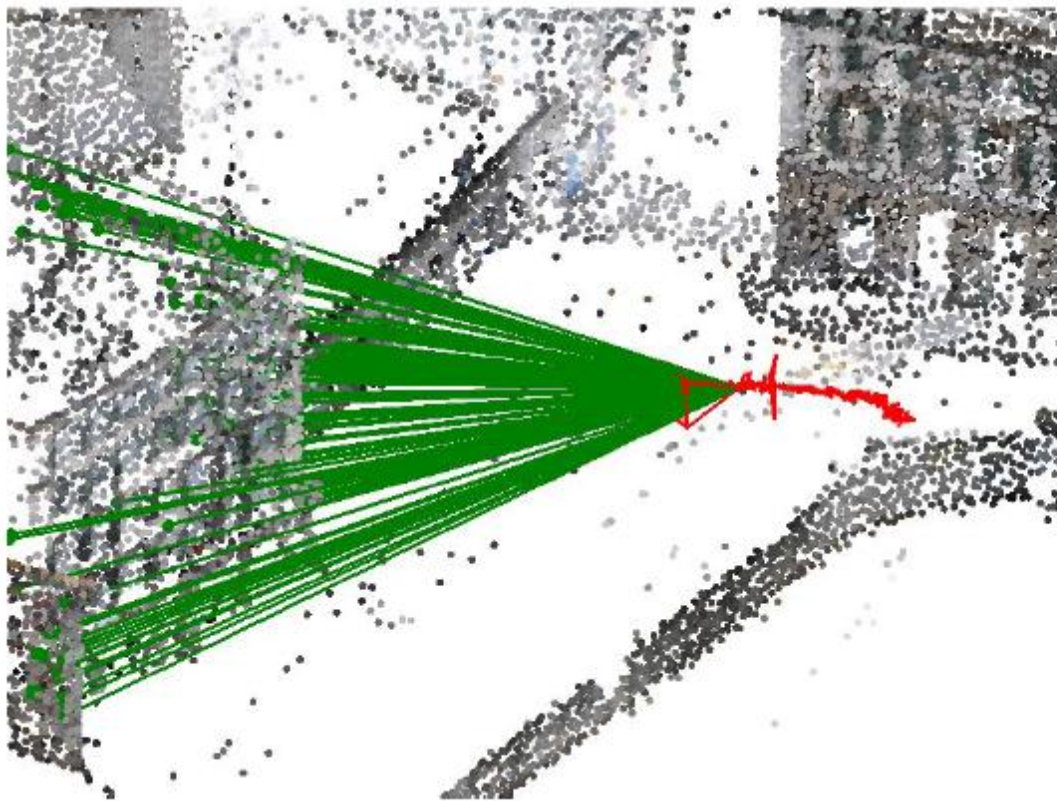


LSD-SLAM: Large-Scale Direct Monocular SLAM,

Jakob Engel and Thomas Schops and Daniel Cremers, ECCV'14

SLAM

- SLAM – технология одновременной реконструкции 3D сцены и оценки положения/параметров движения камеры

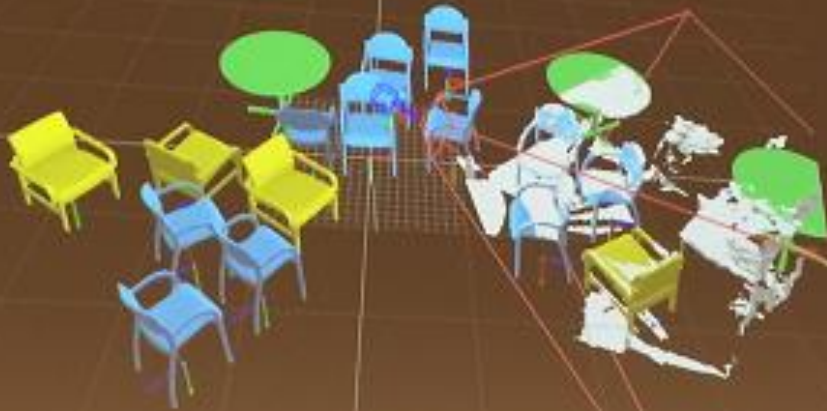


Video Registration to SfM Models,
Till Kroeger and **Luc Van Gool**, ECCV'2014

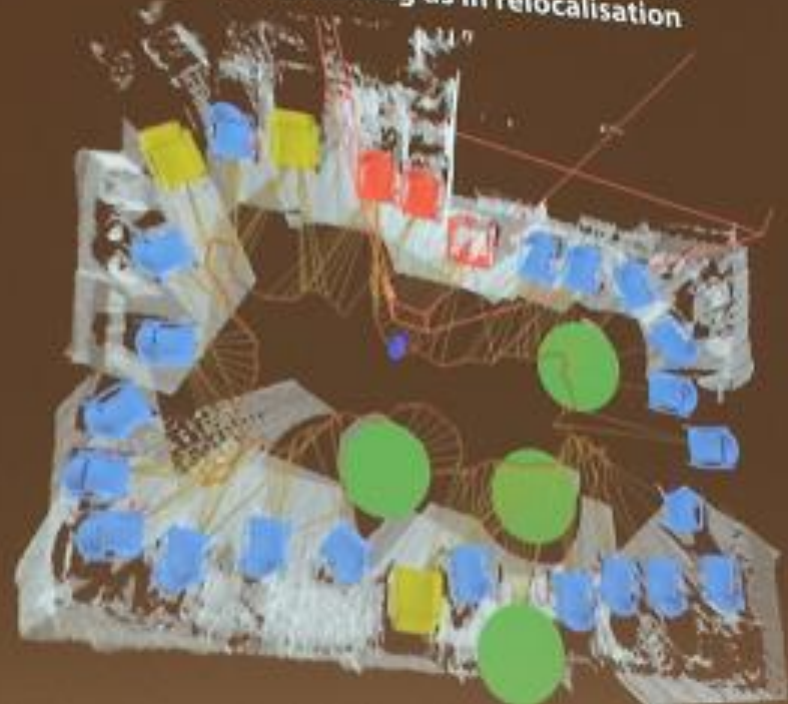
Object-level SLAM



Camera tracking will soon fail
when the user covers the sensor

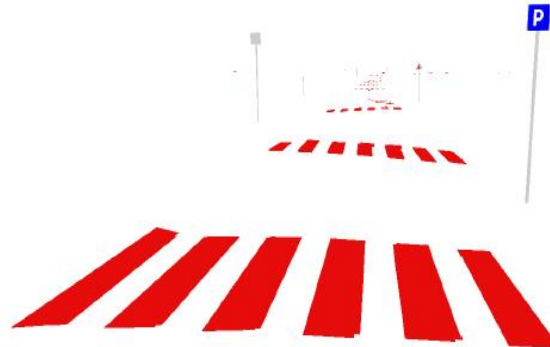
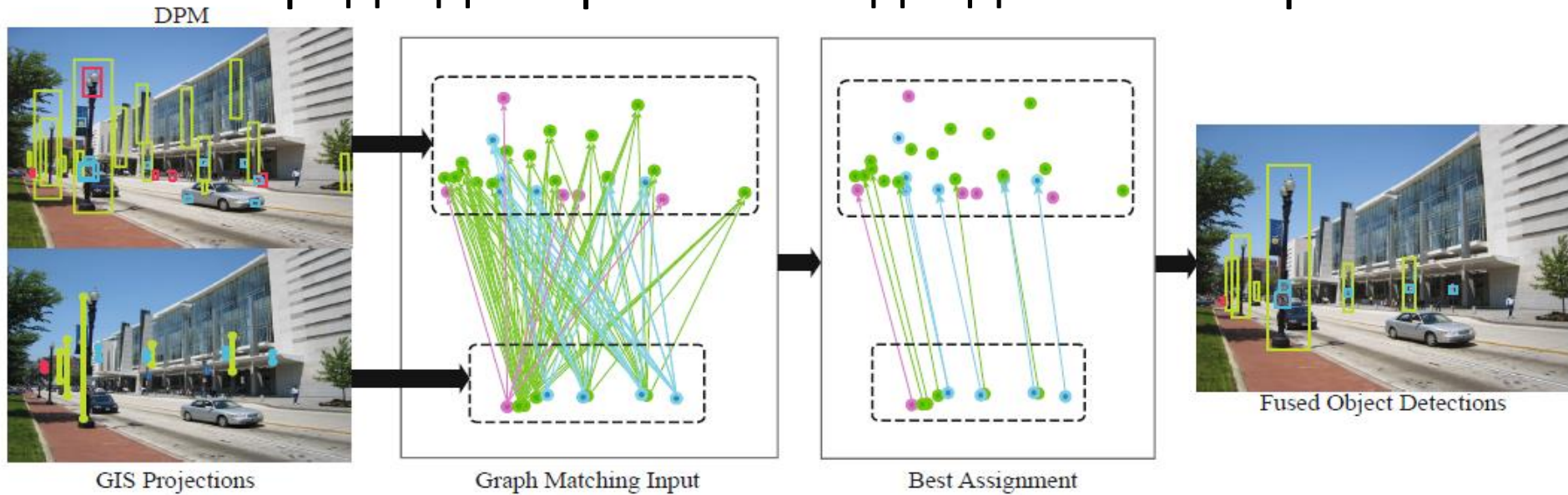


Large loop closures are detected via
graph matching as in relocalisation



Towards real-time, dense tracking, reconstruction and scene understanding,
A. Davison, PCV'14

Распознавание характерных элементов городской среды для привязки видеоданных к карте



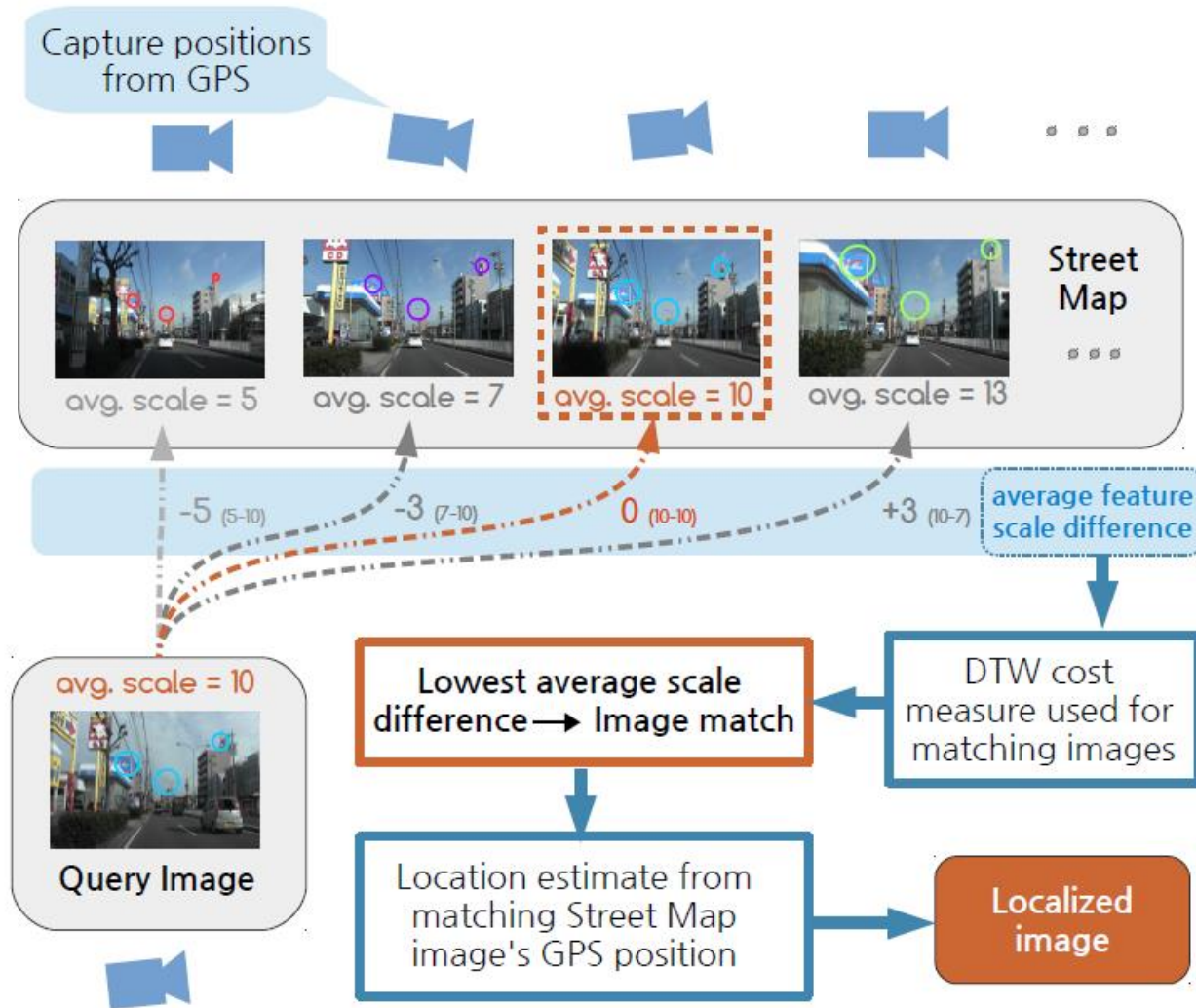
1) *GIS-Assisted Object Detection and Geospatial Localization*,

Shervin Ardeshir, Amir Roshan Zamir, Alejandro Torroella, and **Mubarak Shah**, ECCV'14

2) *Augmenting vehicle localization accuracy with cameras and 3D road infrastructure*

database, Lijun Wei, Bahman Soheilian, Valerie Gouet-Brunet, ECCV'14, W02

Распознавание характерных элементов городской среды для привязки видеоданных к карте



Vision-based Vehicle Localization using a Visual Street Map with Embedded SURF Scale, David Wong, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase, ECCV'14, W02

Распознавание характерных элементов городской среды для привязки видеоданных к карте

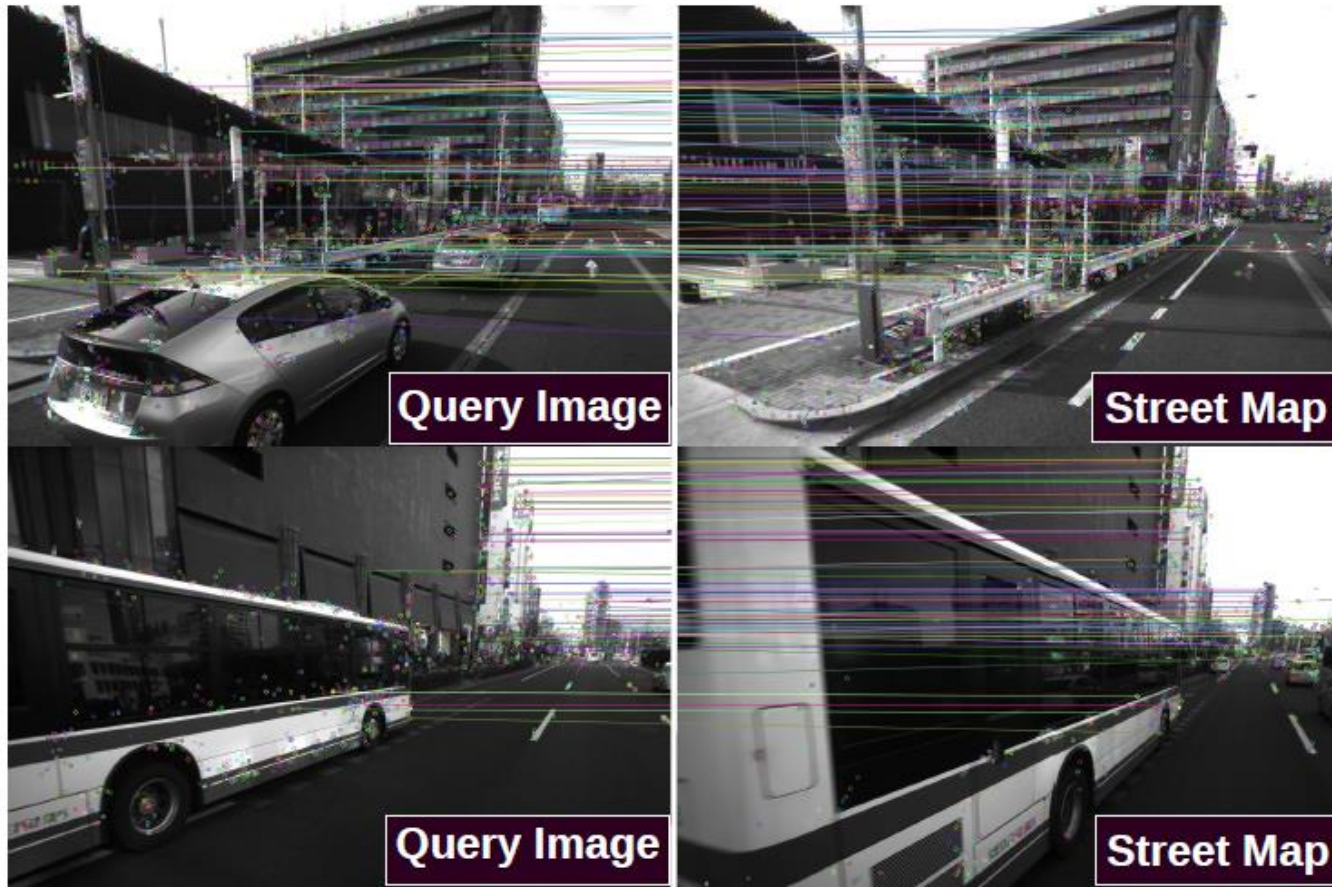


Fig. 5. Example of successful image matching when occlusions occur in either the query image or street map image.

Распознавание характерных элементов городской среды для привязки видеоданных к карте



Deep Features for Text Spotting,

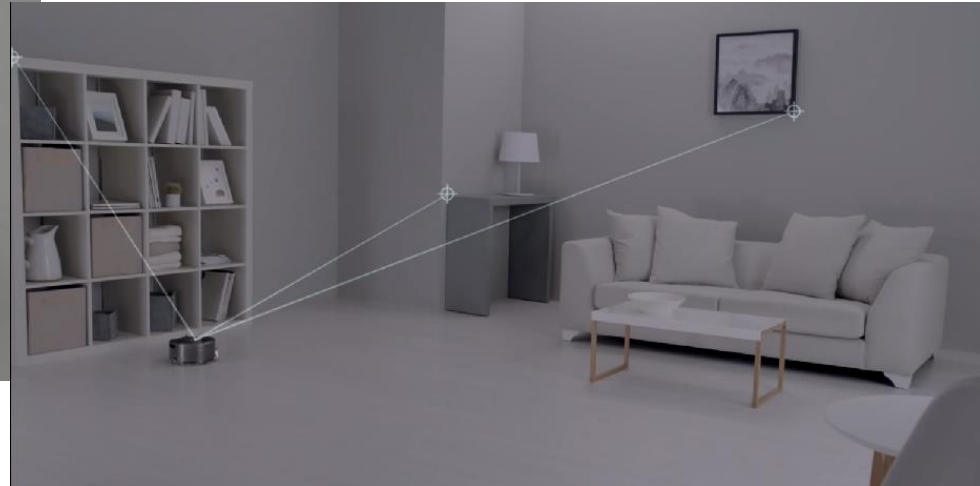
Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman, ECCV'14

Распознавание характерных элементов городской среды для привязки видеоданных к карте



(Huang, Qiao, and Tang, 2014)

SLAM, Autonomous Driving



Towards real-time, dense tracking, reconstruction and scene understanding,
A. Davison, PCV'14

SLAM, UAV, Fish-eye Camera

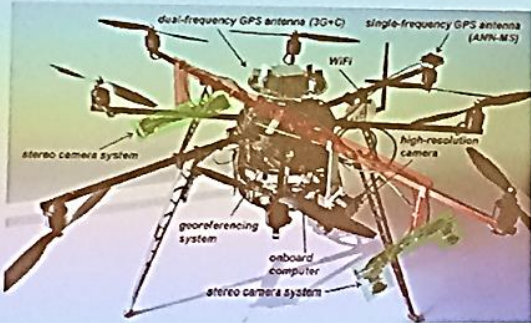


Omnidirectional Visual Odometry

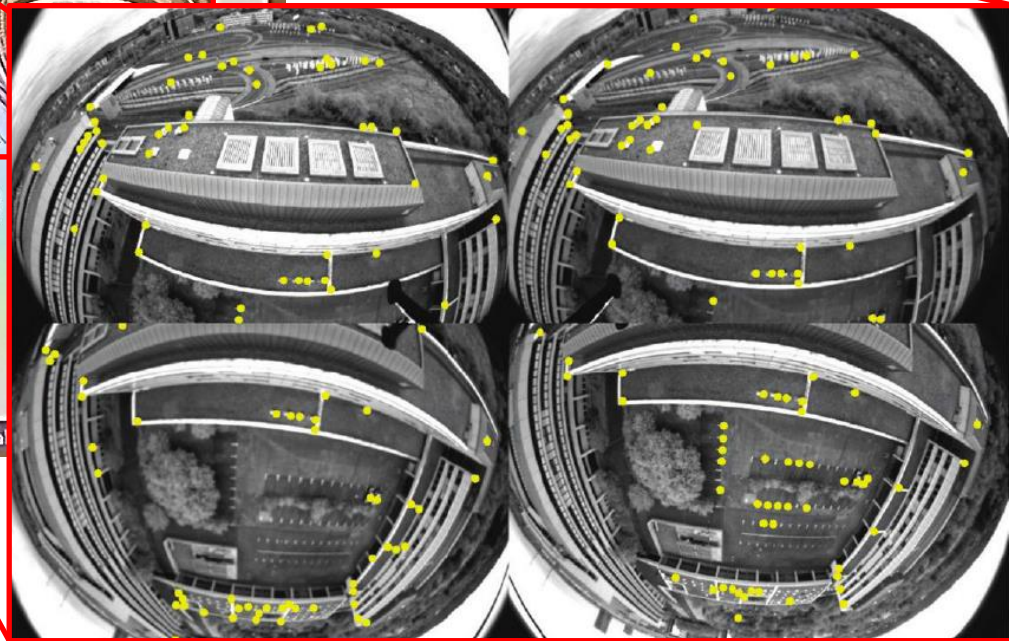


In this talk:

- ▶ Omnidirectional visual odometry system



- ▶ Four image streams (synchronized time of exposure)
 - ▶ Frame rate: 10 Hz
 - ▶ One frame consists of four images
- ▶ RTK-GPS information (direct geo-referencing unit)
 - ▶ accuracy below 3 cm under good conditions



Johannes Schneider Sep 6th, 2014

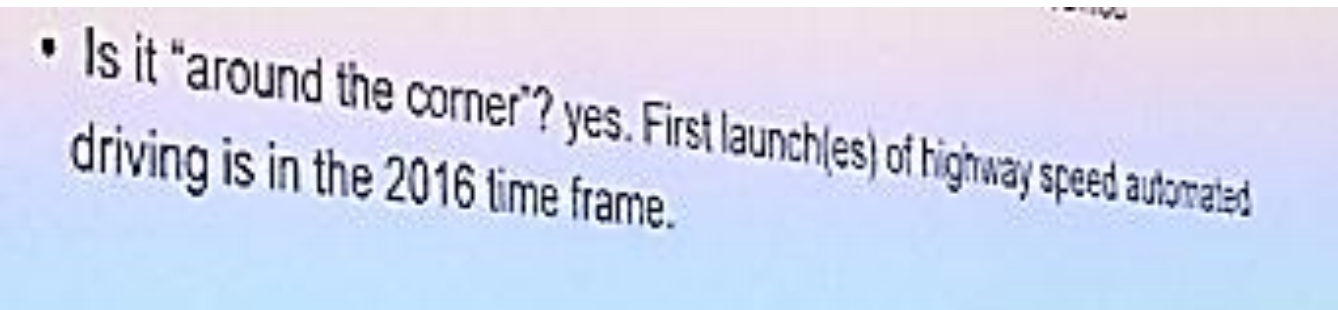
Real

Real-time Accurate Geo-localization of a MAV with Omnidirectional Visual Odometry and GPS, Johannes Schneider and Wolfgang Forstner, ECCV'14, W02

Autonomous Driving

W15 Computer Vision for Road Scene Understanding and Autonomous Driving

- 1400 **Invited Talk:** Is the self-driving car around the corner? Mobileye's work on Computer Vision centric approach to self-driving at consumer level cost, Amnon Shashua, *MobilEye, Israel*
- 1440 **Demo Talk:** Multi-Camera Systems in the V-Charge Project: Fundamental Algorithms, Self Calibration, and Long-Term Localization, *Paul Furgale, ETH Zurich, Switzerland*
- 1500 **Invited Talk:** Intelligent Drive & Pedestrian Safety 2.0, Dariu Gavrila, *Daimler, Germany*



• Is it "around the corner"? yes. First launch(es) of highway speed automated driving is in the 2016 time frame.

Autonomous Driving




Multi-Camera Systems in the V-Charge Project: Fundamental Algorithms, Self Calibration, and Long-Term Localization,
Paul Furgale, ECCV'14, W15



Autonomous Driving

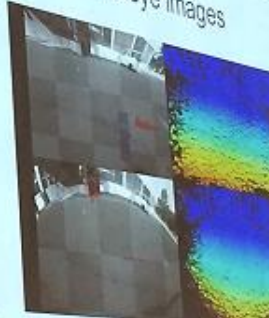
Extraction

- Ground plane extraction
 - Known from extrinsic calibration
- 2D occupancy grid
 - Ground visible, negative weight
 - Obstacle, positive weight
- Obstacle per viewing ray
 - Before obstacle negative weight
 - Behind obstacle positive weight
 - Obstacle where weights fit best




Stereo matching directly on fisheye images

- Larger field-of-view
- Still real-time




Perception | Object Detection and Tracking

Classification based detection - Cylindric images



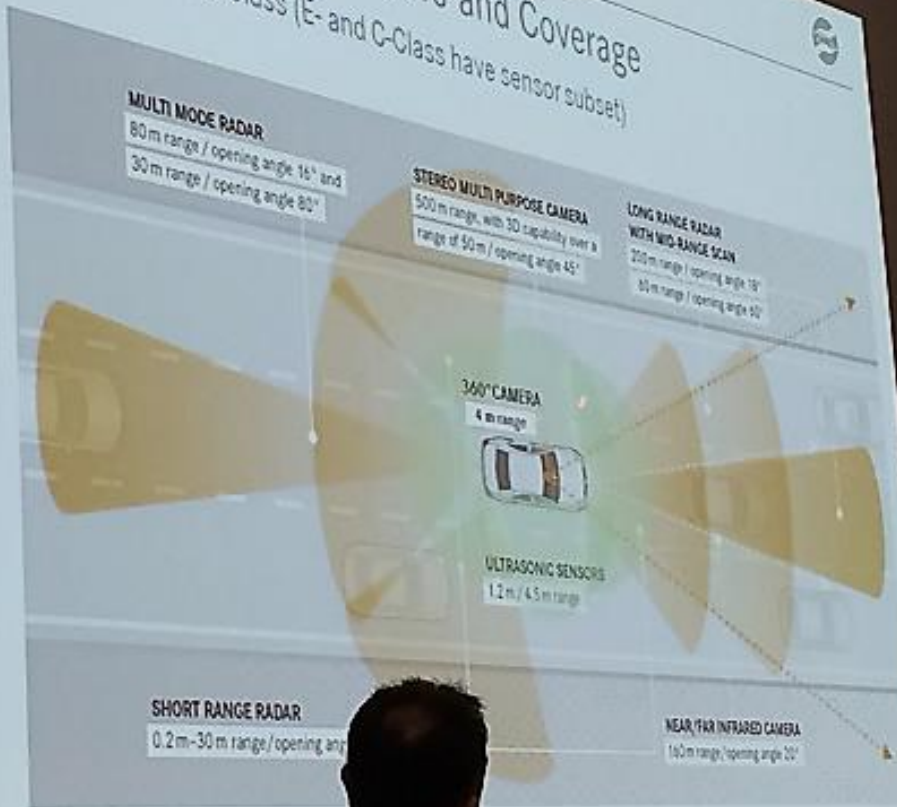
```
Unclap = 1.1, input queue = 1, output queue = 4,
Bucketing = 5, size factor = 12
```



Multi-Camera Systems in the V-Charge Project: Fundamental Algorithms, Self Calibration, and Long-Term Localization,
Paul Furgale, ECCV'14, W15

Autonomous Driving

Intelligent Drive - Sensors and Coverage Mercedes-Benz S-Class (E- and C-Class have sensor subset)



Intelligent Drive - Functionality (MB S-, E- and C-Class, 2013-2014)

Traffic Signs

Adaptive High Beam

Nightview

Attention

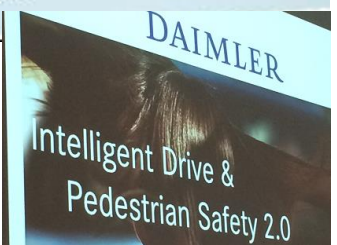
Pre-Crash Braking (longitudinal & lateral traffic) with Pedestrian Recognition

(Active) Body Control

Adaptive Cruise Control with Steering Assist

(Active) Lane Keeping

Intelligent Drive & Pedestrian Safety 2.0,
Darius Gavrila, ECCV'14, W15



Понимание типового

видеосюжета: Multi-Tracking,

Human Detection

Human pose estimation,

Human Action Detection and Prediction,

Crowd behavior, Group analysis

Multi-Tracking

- Системы автоматического анализа специализированных видеоданных (например, некоторых типов спортивных игр)



(Milan, Roth, Schindler, 2014)

Multi-Tracking

- Системы автоматического анализа специализированных видеоданных (например, некоторых типов спортивных игр)



Hybrid Stochastic / Deterministic Optimization for Tracking Sports Players and Pedestrians,
Robert Collins, Penn State University; Peter Carr, Disney Research, ECCV'14

Multi-Tracking

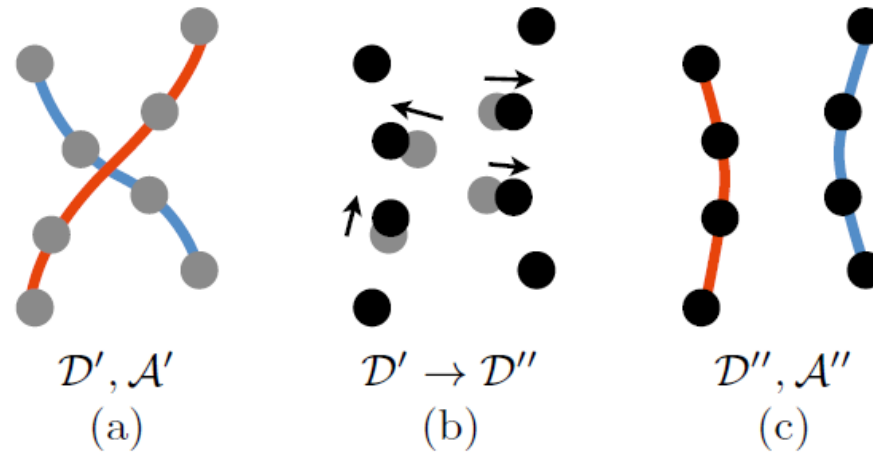


Fig. 1. Stochastic Detection/Deterministic Tracking. (a) An initial set of detections \mathcal{D}' has a corresponding optimal data association solution \mathcal{A}' , shown here as red and blue trajectories. However, due to detection noise, we may have mistakenly swapped the identities of the two targets. (b) If we stochastically perturb the set of detections to generate a new hypothesis \mathcal{D}'' , it may lead (c) to a better data association solution \mathcal{A}'' . Conceptually, we are decomposing the joint optimization of $(\mathcal{D}, \mathcal{A})$ into a stochastic proposal of multi-frame detections \mathcal{D} and a deterministic solution for $\mathcal{A}|\mathcal{D}$ (similar to ‘line search’) given each such proposal.

(Collins, Carr, 2014)

Multi-Tracking

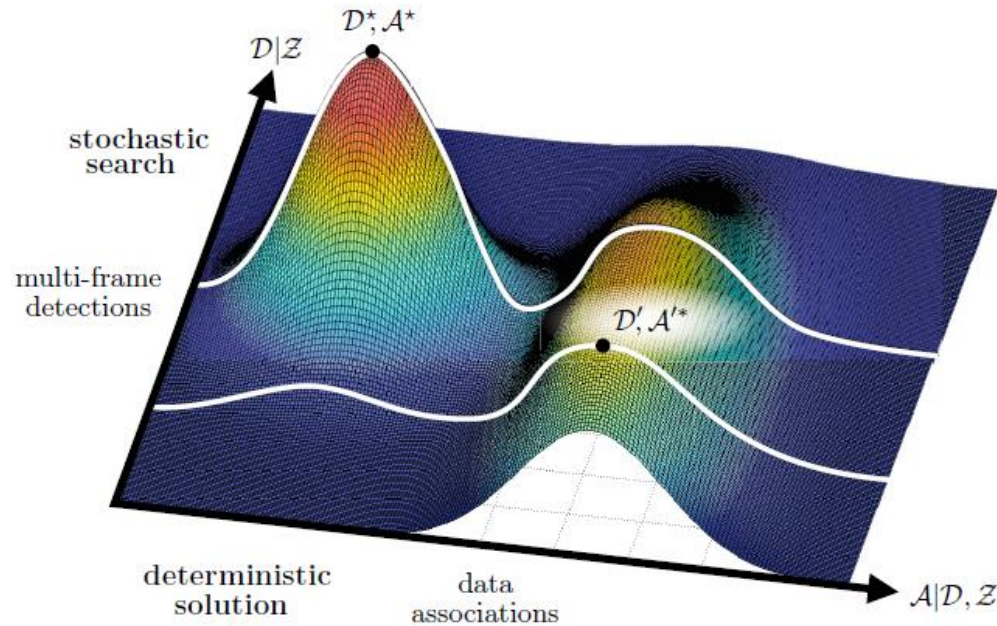


Fig. 2. Hybrid Stochastic/Deterministic Optimization. The goal is to determine the optimal set of detections \mathcal{D}^* and associations \mathcal{A}^* for observations \mathcal{Z} . We factor the joint optimization into stochastic search over detections $\mathcal{D}|\mathcal{Z}$ interleaved with deterministic solutions for associations $\mathcal{A}|\mathcal{D}, \mathcal{Z}$. Each hypothesized set of detections \mathcal{D}' results in a reduced ‘line search’ for the corresponding best set of associations \mathcal{A}' (which has a deterministic solution for energy functions of pairwise potentials).

(Collins, Carr, 2014)

Multi-Tracking

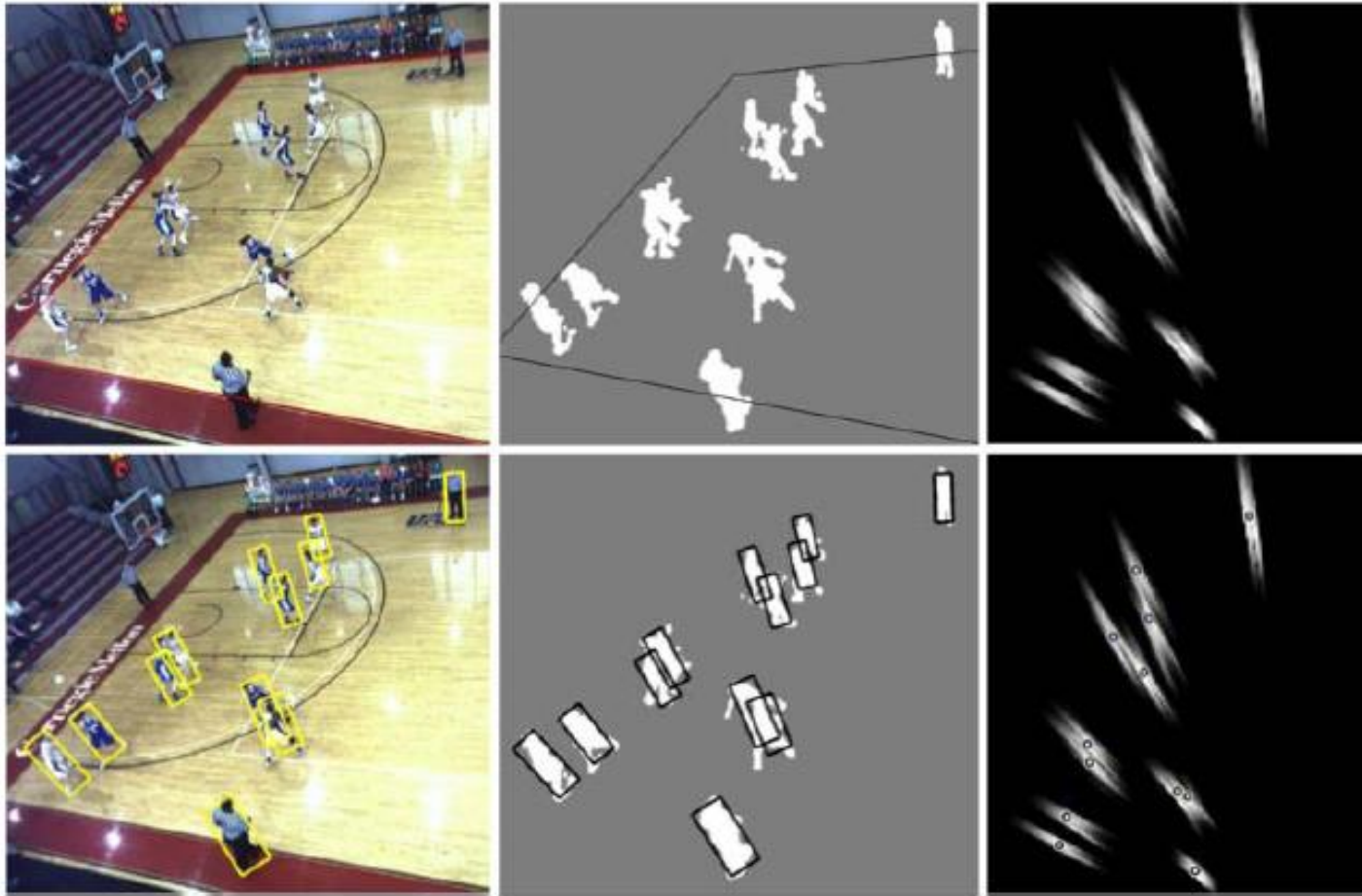


Fig. 3. Top row, left to right: color image I_k ; foreground mask F_k (also showing region of interest); ground plane proposal map M_k . Bottom row: single frame detection results overlaid on each form of observation data.

(Collins, Carr, 2014)

Multi-Tracking, Fusion

Overview

Observation:

- Different object tracking algorithms have different, sometimes even contrary strengths.
- Practically, even outdated methods can sometimes clearly outperform state-of-the-art methods.
Example: SMS performs much better than SCM on the lemming sequence. In average it is contrary. See figures.

Idea:

- Create a fusion approach that outperforms tracking algorithms by fusing their results in a suitable way.

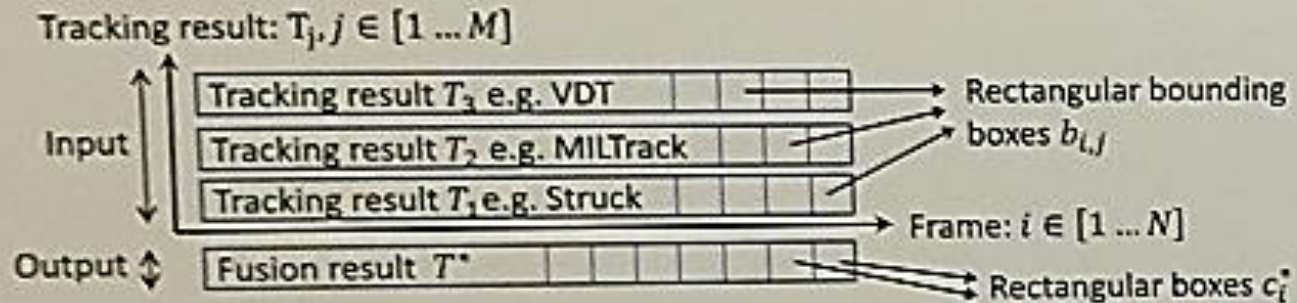
The Fusion Approach:

- Our fusion approach performs considerably better than the best tracking algorithms used for fusion.
- The approach is very generic and only needs tracking results in the form of rectangular boxes as input.
- We introduce a basic approach and extensions like dynamic programming based trajectory optimization.
- Online fusion is possible. With trajectory optimization most discontinuities can be avoided - even online.
- Our fusion approach often even outperforms the best tracking algorithm on a sequence (by up to 33%).
- Probably, our approach is also able to outperform future tracking algorithms by fusing their results.

Multi-Tracking, Fusion

Details of our Approach

Notation



Basic Approach

Idea: Use attraction fields between tracking results and find the position of maximum attraction. Attraction does not require sequence dependent threshold parameters like our previous work [1]. The attraction energy for a candidate box c in a frame i is calculated as:

$$a_i(c) = \sum_{j \in M} \frac{w_j^2}{d(b_{i,j}, c)^2 + \sigma} \quad \leftarrow w_j = 1 \text{ for the basic approach}$$

The distance $d(b_{i,j}, c)$ is calculated in the 4 dimensional (x,y,width,height) space (see paper). σ is a constant that is useful for noise reduction. The (nearly final) fusion result is: $c_i^* = \underset{c \in b_{i,1} - b_{i,M}}{\operatorname{argmax}} (a_i(c))$.

Tracker Weights

Idea: Trust algorithms more that perform on average better by weighting them. Let G_s^i be the ground truth labeling for a sequence s at frame i . Then the weight w_j is determined as:

$$w_j = \sum_{s \in S} \sum_{i \in N} \frac{1}{d(G_s^i, b_{i,j}^s)^2 + \sigma}$$

Multi-Tracking, Fusion

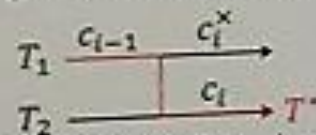
Trajectory Optimization

Idea: Do not treat frames independently. Instead find an optimal fusion result for the whole trajectory to avoid discontinuities in the fusion result. Energy function E_T for the whole trajectory T :

$$E_T = \sum_{i \in N} \overline{a}_i(c_i) + \beta p(c_{i-1}, c_i), \quad c_i \in R_i = \{b_{i,1} \dots b_{i,M}\}, \quad T = \{c_1 \dots c_N\}$$

$\overline{a}_i(c) = a_i(c) / \max_c(a_i(c))$ is normalized to consider frames equally in E_T . β weights the importance of continuity. p penalizes tracking results switches from one frame to the next:

$$p(c_{i-1}, c_i) = \frac{\sigma}{d(c_i^x, c_i)^2 + \sigma}, \quad c_{i-1} = b_{i-1,j} \leftrightarrow c_i^x = b_{i,j}$$



The trajectory T^* that maximizes E_T can be efficiently calculated with dynamic programming, by calculating the following $N \times M$ energy fields in increasing frame order:

$$E(0, j) = \overline{a}_i(b_{0,j}) \quad E(i, j) = \overline{a}_i(b_{i,j}) + \max_{j_2 \in M} p(b_{i-1,j_2}, b_{i,j}) + E(i-1, j_2)$$

Online fusion result: $c_i^* = \max_{j=1 \dots M} E(i, j)$. For offline fusion a lookup table is required, see paper.

Tracking algorithm removal

Idea: Remove bad tracking results before fusion as they may disturb fusion more than they support it.

Local Approach:

- Removes different results for each sequence.
- Idea: Find tracking results that are likely least useful for fusion by calculating the overall attraction each result gets in a sequence: $P_j = \sum_{i \in N} a_i^w(b_{i,j})$. Remove γ results with the worst P_j before fusion.

Global Approach:

- Removes tracking algorithms permanently
- Idea: Divide the training dataset into 10 overlapping sets that contain 90% of the dataset, each. If for at least 7 sets fusion becomes better by removing an algorithm remove it permanently.

Multi-Tracking, Fusion

- Performance is measured in average overlap. Our fusion approaches outperform the best tracking algorithm on a sequence on 11,15,20,18 and 22 sequences respectively. They gain at least 95% of the performance of the best algorithm on a sequence on 25,27,33,35 and 34 sequences, respectively.

Success plots for OPE (and SRE,TRE). Our fusion approaches clearly outperform the best tracking algorithms as well as our previous work. The figures show our fusion approaches, the 5 best and 2 worst tracking algorithms and the average of all 29 tracker curves. The gray curves are theoretical bounds.

Runtime performance

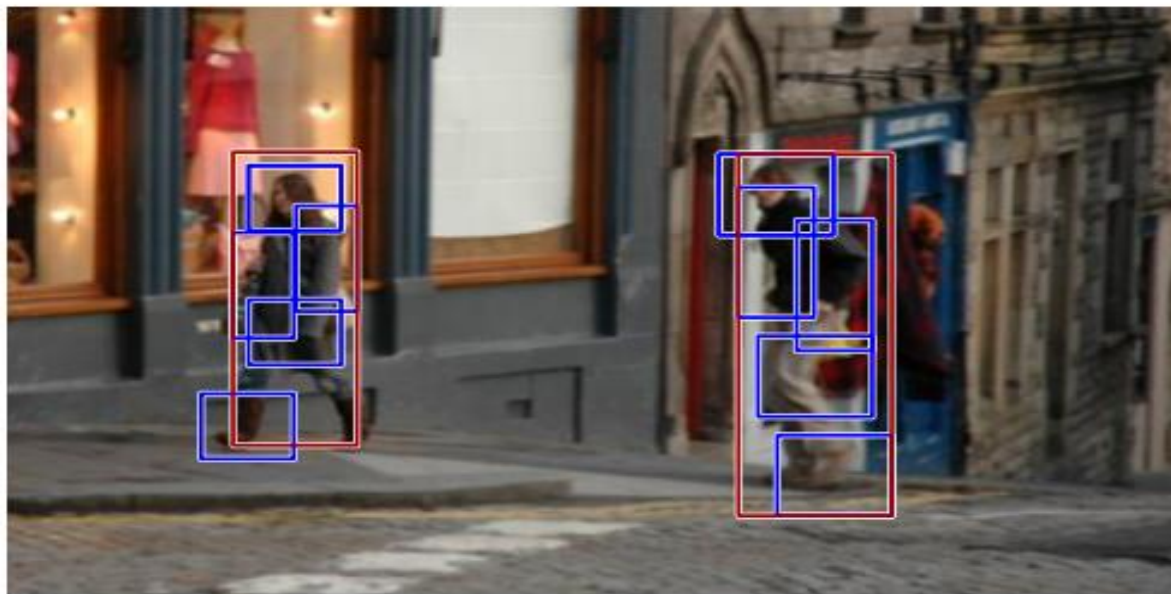
Overall runtime of the fusion based tracking approach for different tracking algorithm selections (see paper).

With 9 frames/sec. we can already outperform the best algorithm SCM.

Our trajectory optimization approaches perform well here. The online version is closer to the ground truth than most tracking algorithms. The offline versions perform even better.

The best algorithm SCM (score: 0.5) can already be outperformed with the 15 worst algorithms which have a score of only 0.21 - 0.36.

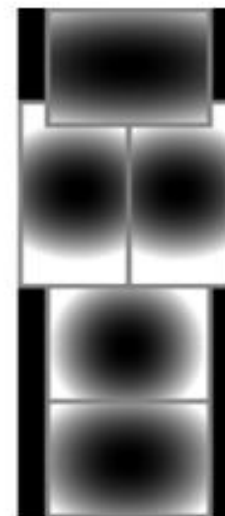
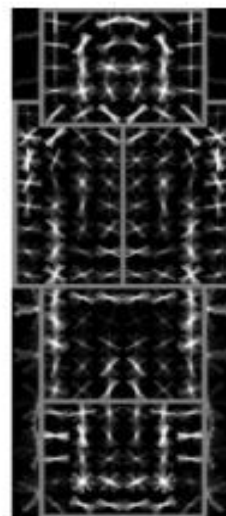
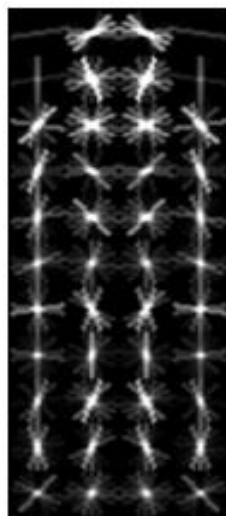
Обнаружение людей (пешеходов)



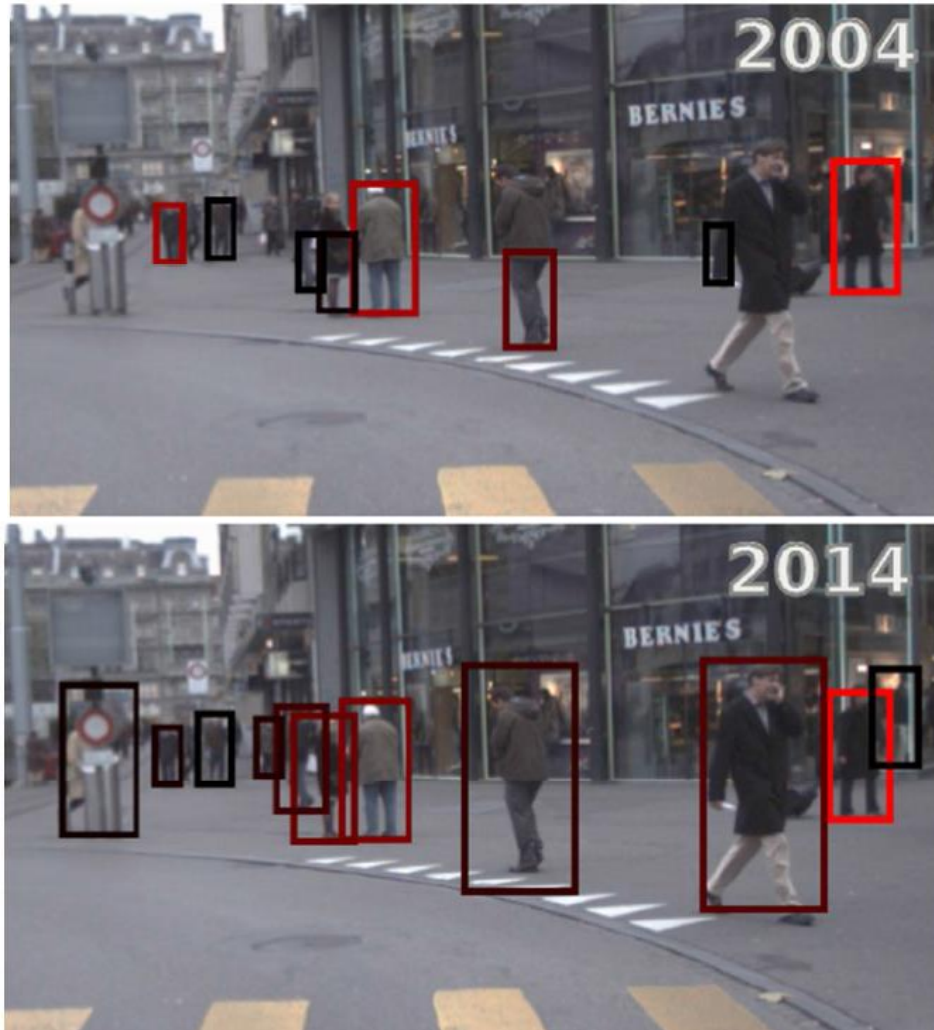
State-of-Art:

(Felzenszwalb, Girshick, McAllester and Ramanan, 2010)

(Dollár, Wojek, Schiele, Perona, 2012)



Обнаружение людей (пешеходов)



Ten years of pedestrian detection, what have we learned?

R. Benenson. et. Al, ECCV 2014

Обнаружение людей (пешеходов)

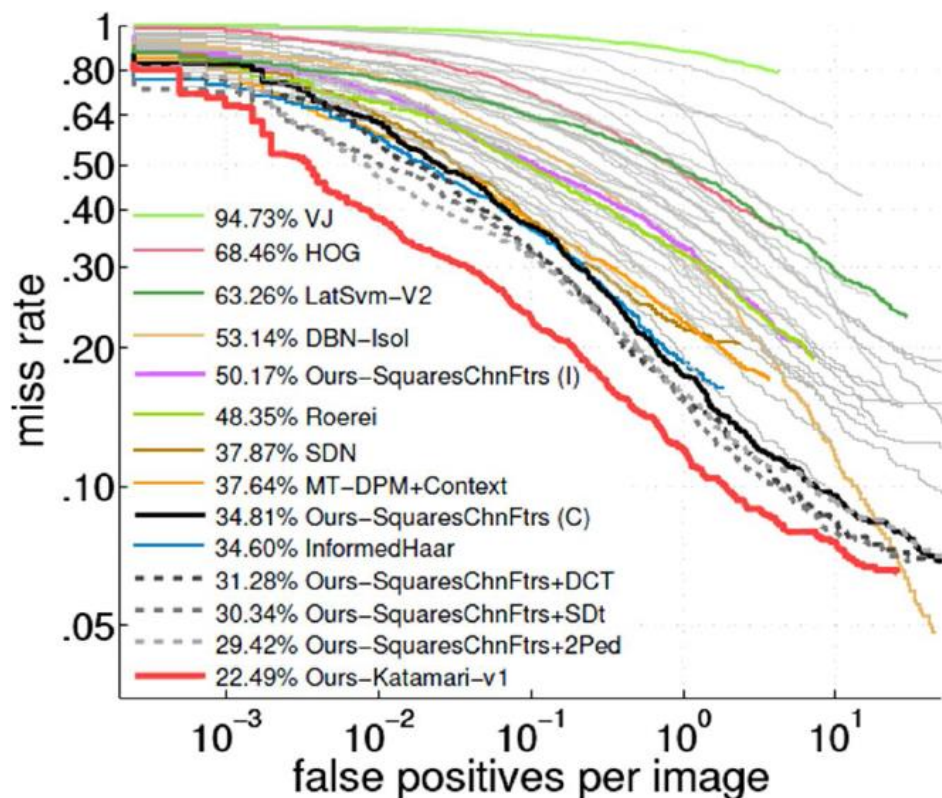


Figure 7: Some of the top quality detection methods for Caltech-USA. See section [4.2](#).

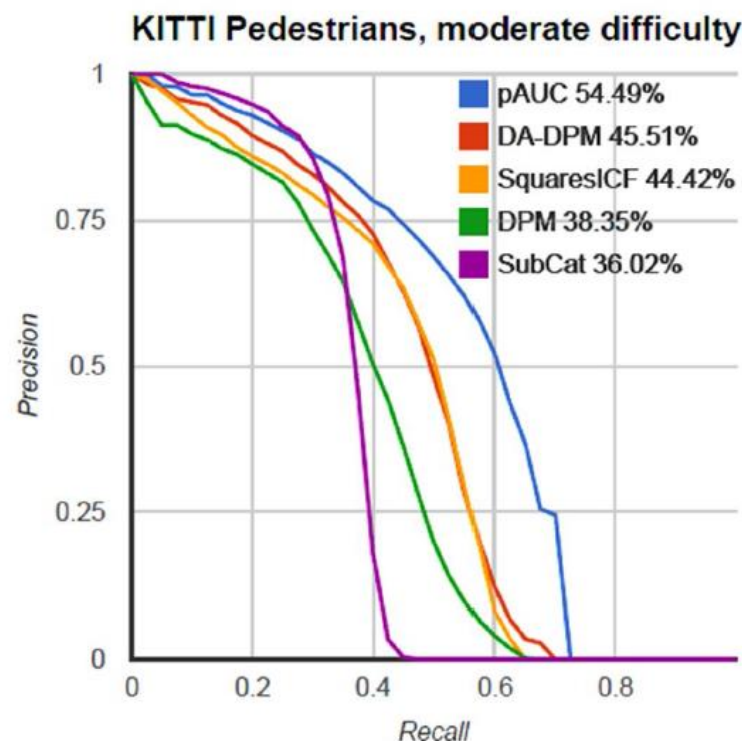


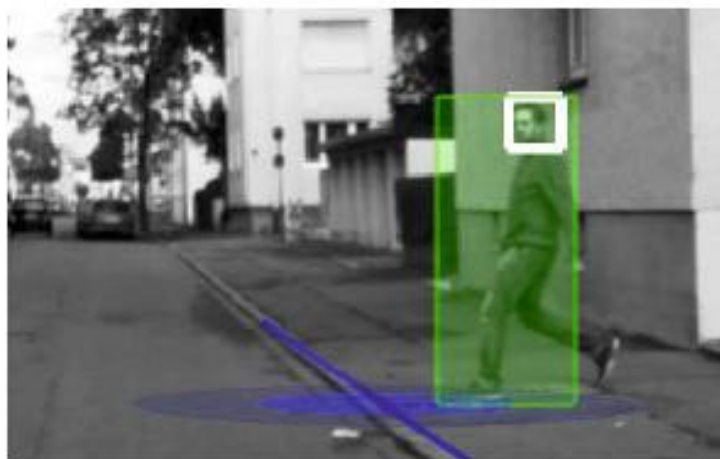
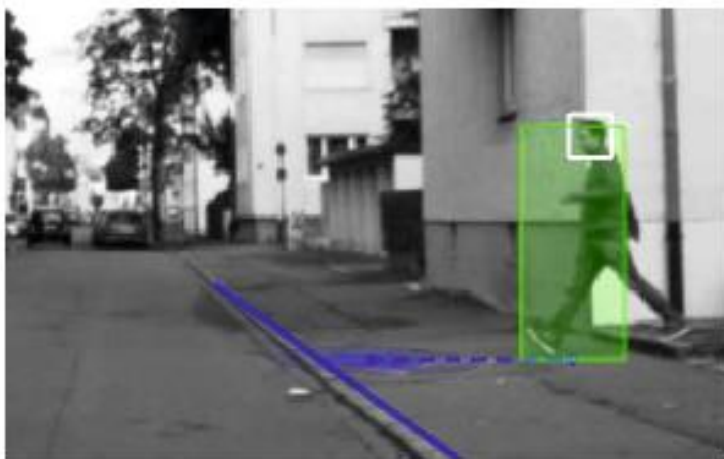
Figure 8: Pedestrian detection on the KITTI dataset.

Ten years of pedestrian detection, what have we learned?

R. Benenson. et. Al, ECCV 2014

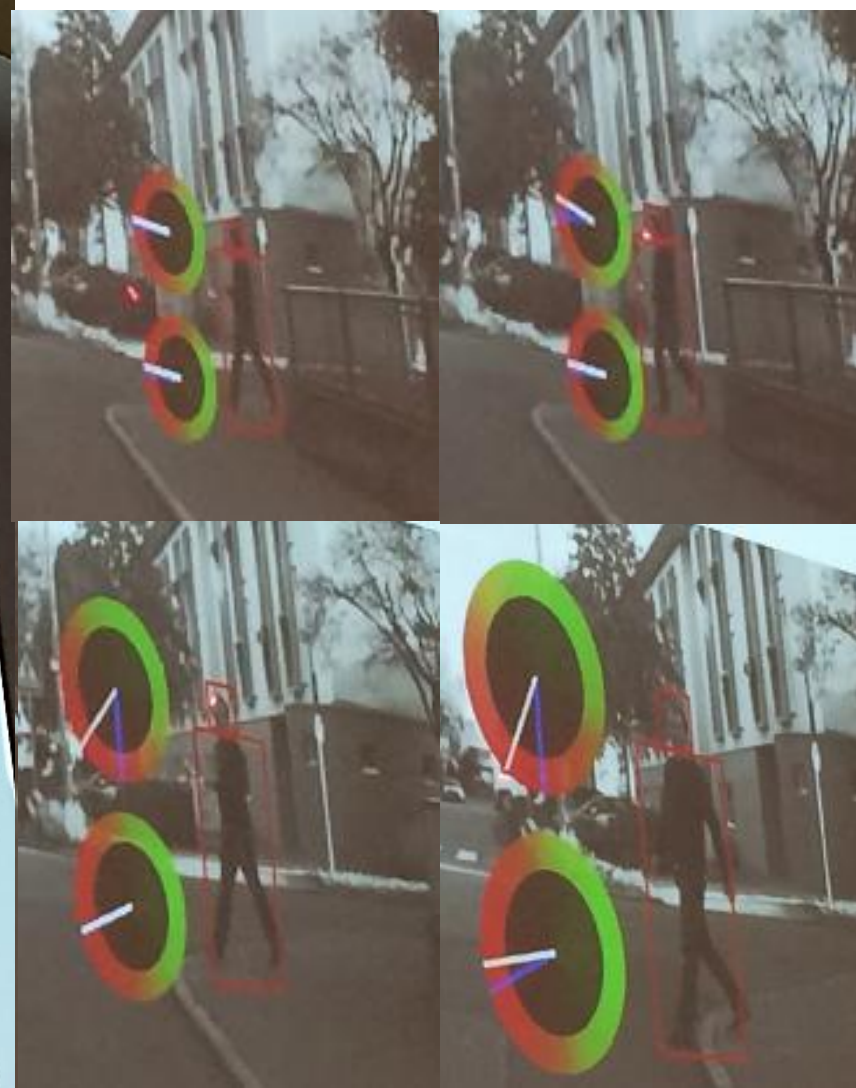
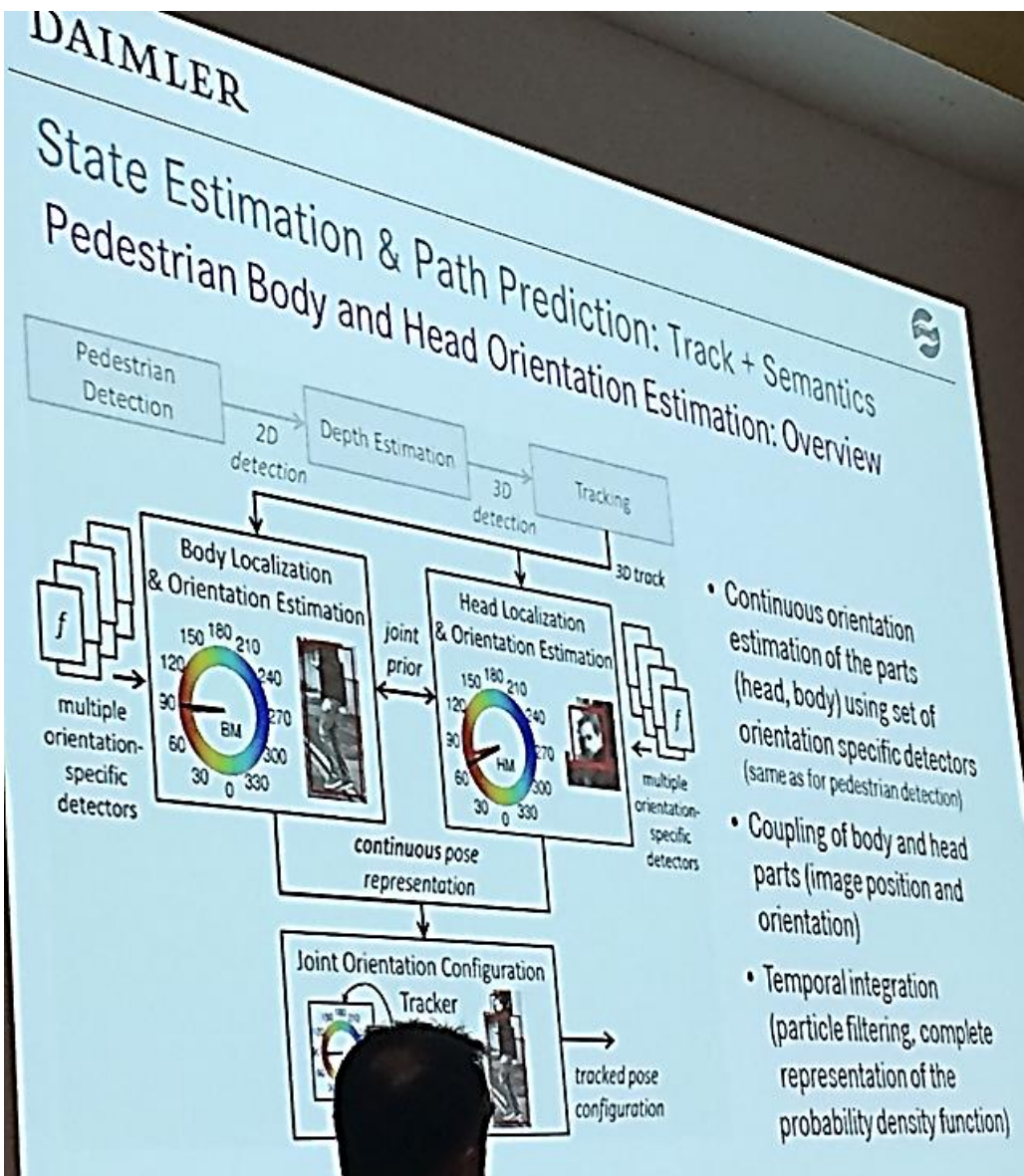
Предсказание поведения людей (пешеходов)

- Предсказание поведения людей (пешеходов):
Перейдет ли пешеход дорогу?



Context-Based Pedestrian Path Prediction,
Kooij, Schneider, Flohr, and **Gavrila**, ECCV'14

Предсказание поведения людей (пешеходов)



Context-Based Pedestrian Path Prediction, Kooij, Schneider, Flohr, and Gavrila, ECCV'14

Crowd behavior, Group analysis

- Оценка характера поведения групп людей или толпы
- Выделение и прослеживание отдельных людей в толпе



Ре-идентификация людей при съемке в различных условиях



VIPeR dataset



PRID dataset

(T. Wang et al., 2014)

Ре-идентификация людей при съемке в различных условиях

- Гистограммы цветовых и геометрических свойств



Salient Color Names for Person Re-identification,

Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li¹, ECCV'14

Ре-идентификация людей при съемке в различных условиях



(a) Cross-view lighting variations



(b) Camera viewpoint changes

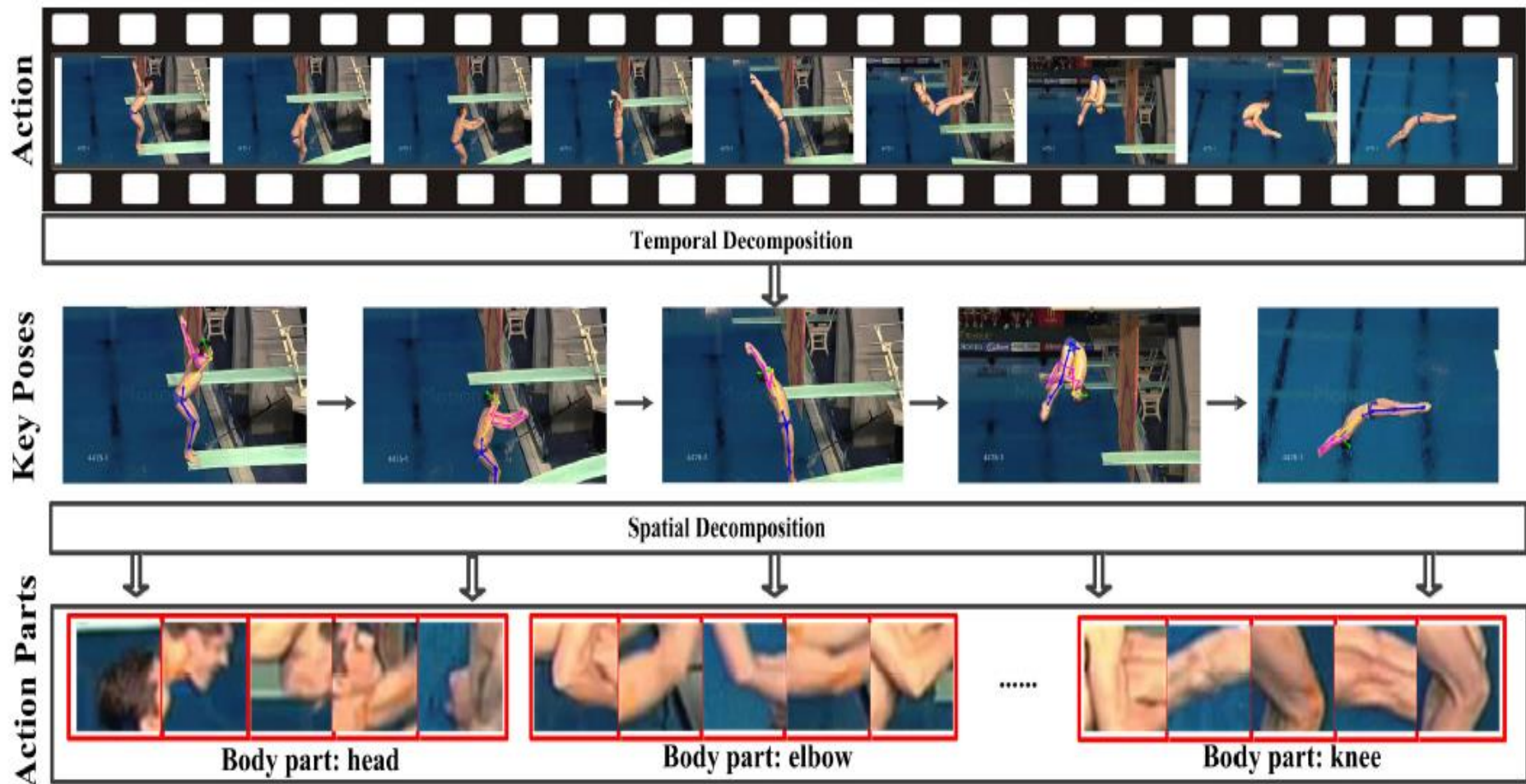


(c) Clothing similarity



(d) Background clutter & occlusions

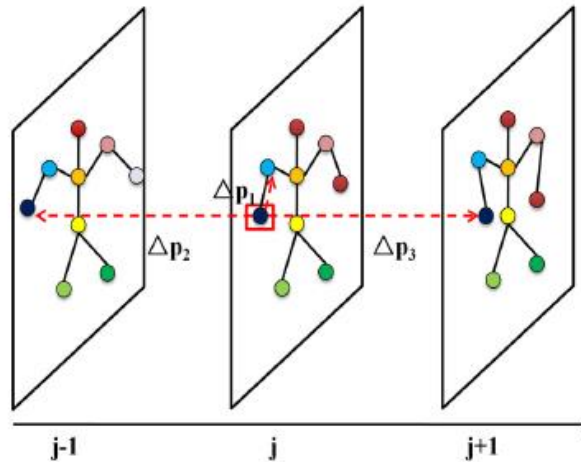
Распознавание действий людей



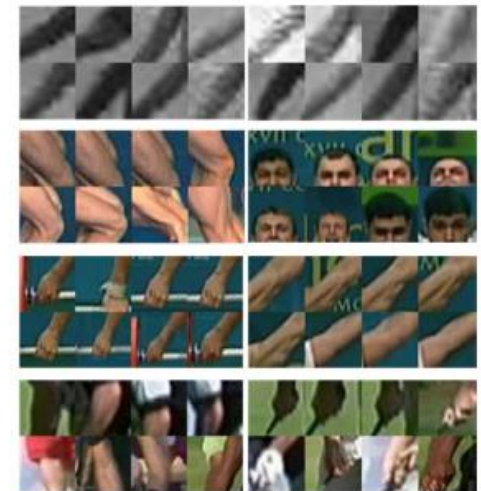
Распознавание действий людей



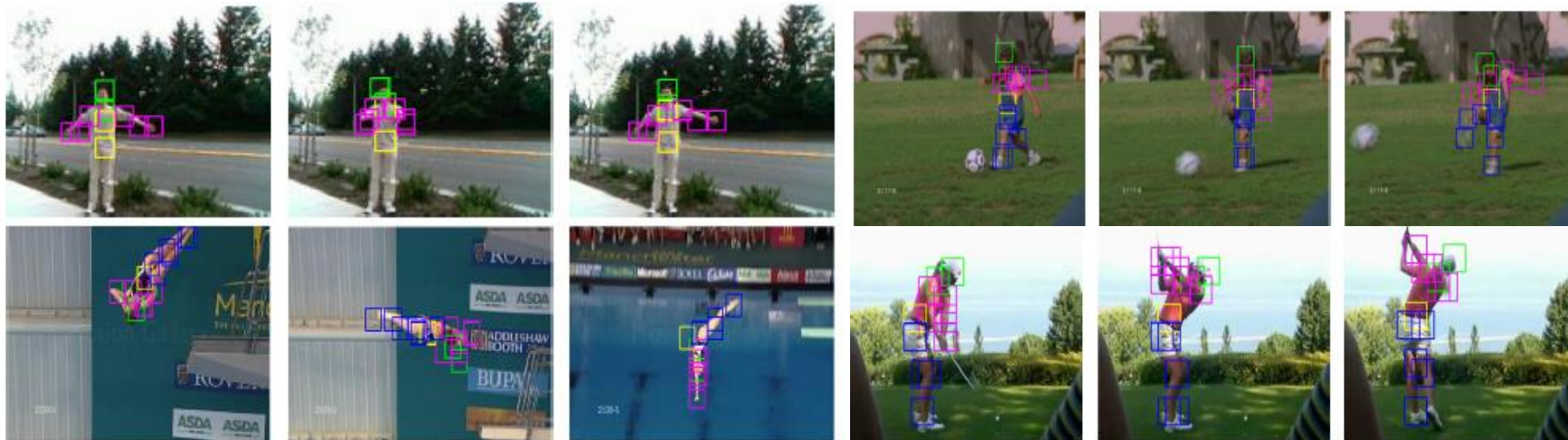
(a) examples of annotations



(b) calculation of descriptor



(c) dynamic-poselets



Распознавание взаимодействий людей



UT-Interaction 1

UT-Interaction 2



RGB

Depth

Skeleton

Gnd truth

Observation

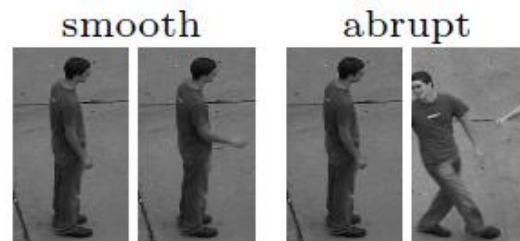
Simulation



likely

unlikely

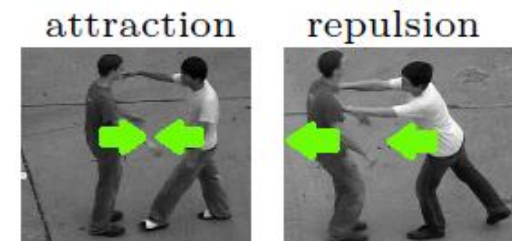
Cooccurrence



smooth

abrupt

Transition



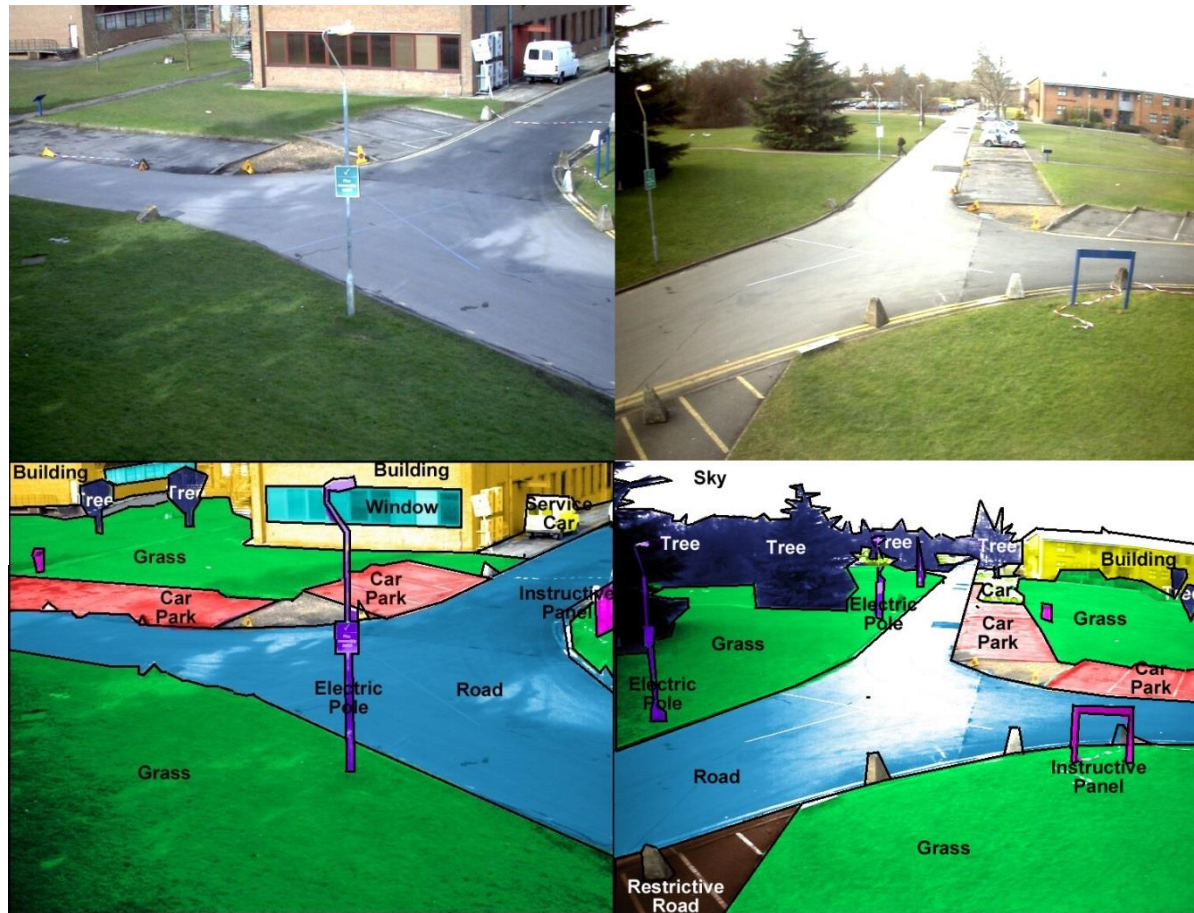
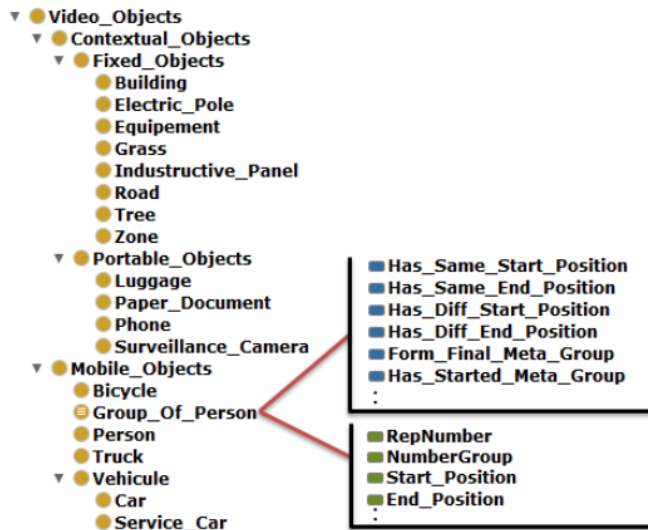
attraction

repulsion

Symmetry

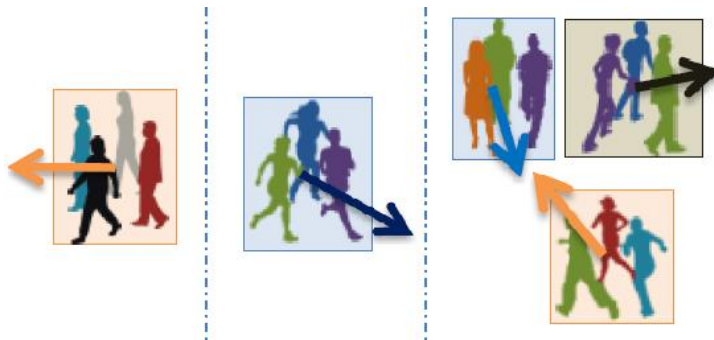
Action-Reaction: Forecasting the Dynamics of Human Interaction,
De-An Huang and Kris M. Kitani, ECCV'14

Построение и использование пространственно-временных логик и онтологий для анализа сложных динамических сцен



Events detection using a video-surveillance Ontology and a rule-based approach,
Yassine Kazi Tani, Adel Lablack, Abdelghani Ghomari, and Ioan Marius Bilasco, ECCV'14

Построение и использование пространственно-временных логик и онтологий для анализа сложных динамических сцен



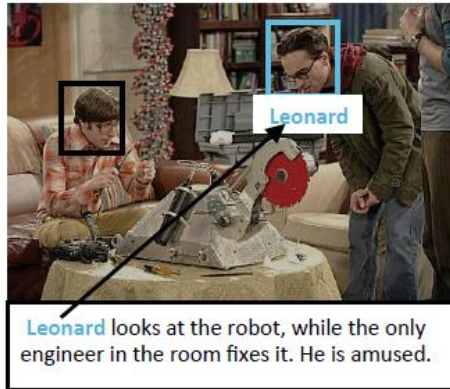
Group walking, Group running, Group merging and Group splitting.



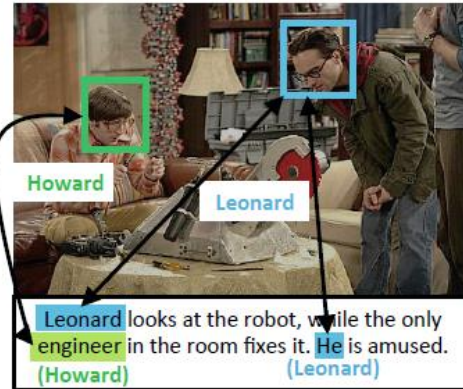
```
BB(?BBx), BB(?BBy), Frame(?F1), MBB(?MBB1), MBB(?MBB2), BB_Detected_In_Frame(?BBx, ?F1), BB_Detected_In_Frame(?BBy, ?F1), BB_Bottom_Left_Point_Y(?BBx, ?h), BB_Bottom_Right_Point_Y(?BBy, ?d), BB_Number(?BBx, ?n4), BB_Number(?BBy, ?n5), BB_Top_Left_Point_X(?BBx, ?a), BB_Top_Left_Point_X(?BBy, ?f), BB_Top_Left_Point_Y(?BBx, ?e), BB_Top_Left_Point_Y(?BBy, ?i), BB_Top_Right_Point_X(?BBx, ?j), BB_Top_Right_Point_X(?BBy, ?b), BB_Top_Right_Point_Y(?BBy, ?c), MBB_ID(?MBB1, ?n1), MBB_ID(?MBB2, ?n1), Number_BB_In_Frame(?F1, 2), Number_Frame(?F1, ?n1), Number_MBB(?MBB1, ?n2), Number_MBB(?MBB2, ?n3), add(?x2, ?b, 20), greaterThan(?a, ?b), greaterThan(?h, ?d), greaterThan(?n3, ?n2), greaterThanOrEqual(?b, ?x1), greaterThanOrEqual(?e, ?c), lessThanOrEqual(?a, ?x2), lessThanOrEqual(?e, ?d), subtract(?x1, ?a, 20), subtract(?z1, ?, ?f), subtract(?z2, ?h, ?i) -> BB_Represent_MBB(?BBx, ?MBB1), BB_Represent_MBB(?BBy, ?MBB1), MBB_Detected_In_Frame(?MBB1, ?F1), MBB_H(?MBB1, ?z1), MBB_Top_Left_Point_X(?MBB1, ?f), MBB_Top_Left_Point_Y(?MBB1, ?i), MBB_W(?MBB1, ?z2)
```

Events detection using a video-surveillance Ontology and a rule-based approach,
Yassine Kazi Tani, Adel Lablack, Abdelghani Ghomari, and Ioan Marius Bilasco, ECCV'14

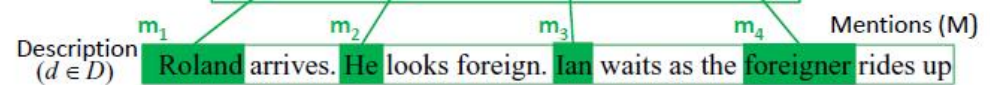
Автоматическое аннотирование видеоданных с использованием текстовых тегов



(a) One directional model



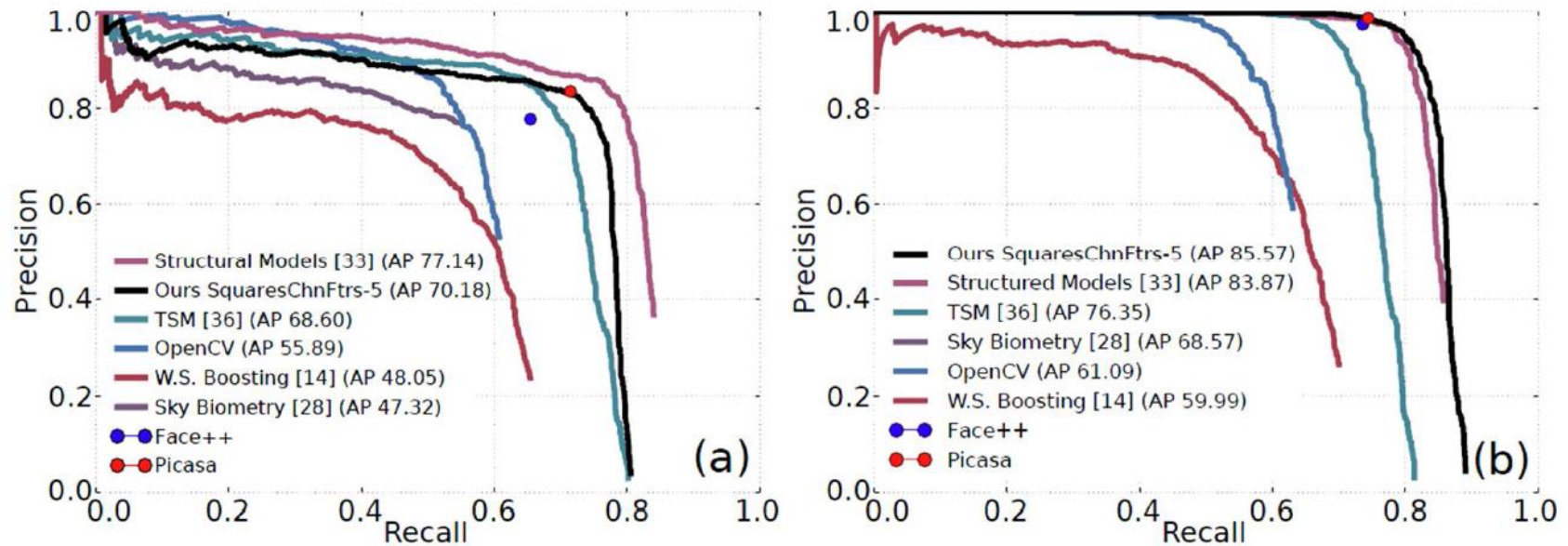
(b) Bidirectional model



Linking People in Videos with “Their” Names Using Coreference Resolution, Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei, ECCV’14

Face Detection and Recognition

Обнаружение лиц

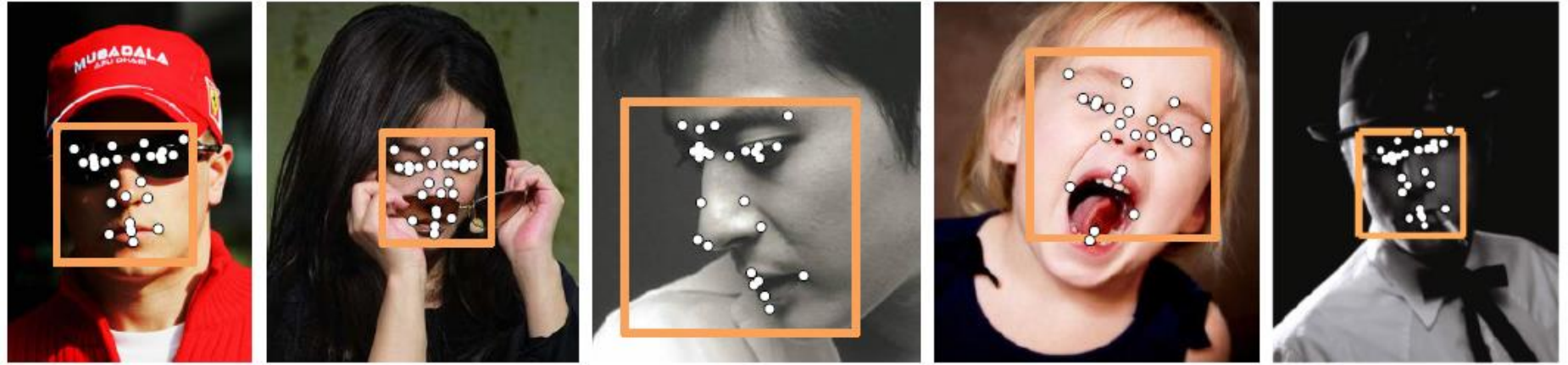


Не в реальном времени: по оценкам, при использовании всех ускорений, на GPU или многоядерной системе может достичь 10 кадров/сек

Face detection without bells and whistles.

M. Mathias. et. alECCV2014

Распознавание лиц в сложных условиях съемки, при низком разрешении, при наличии мимики



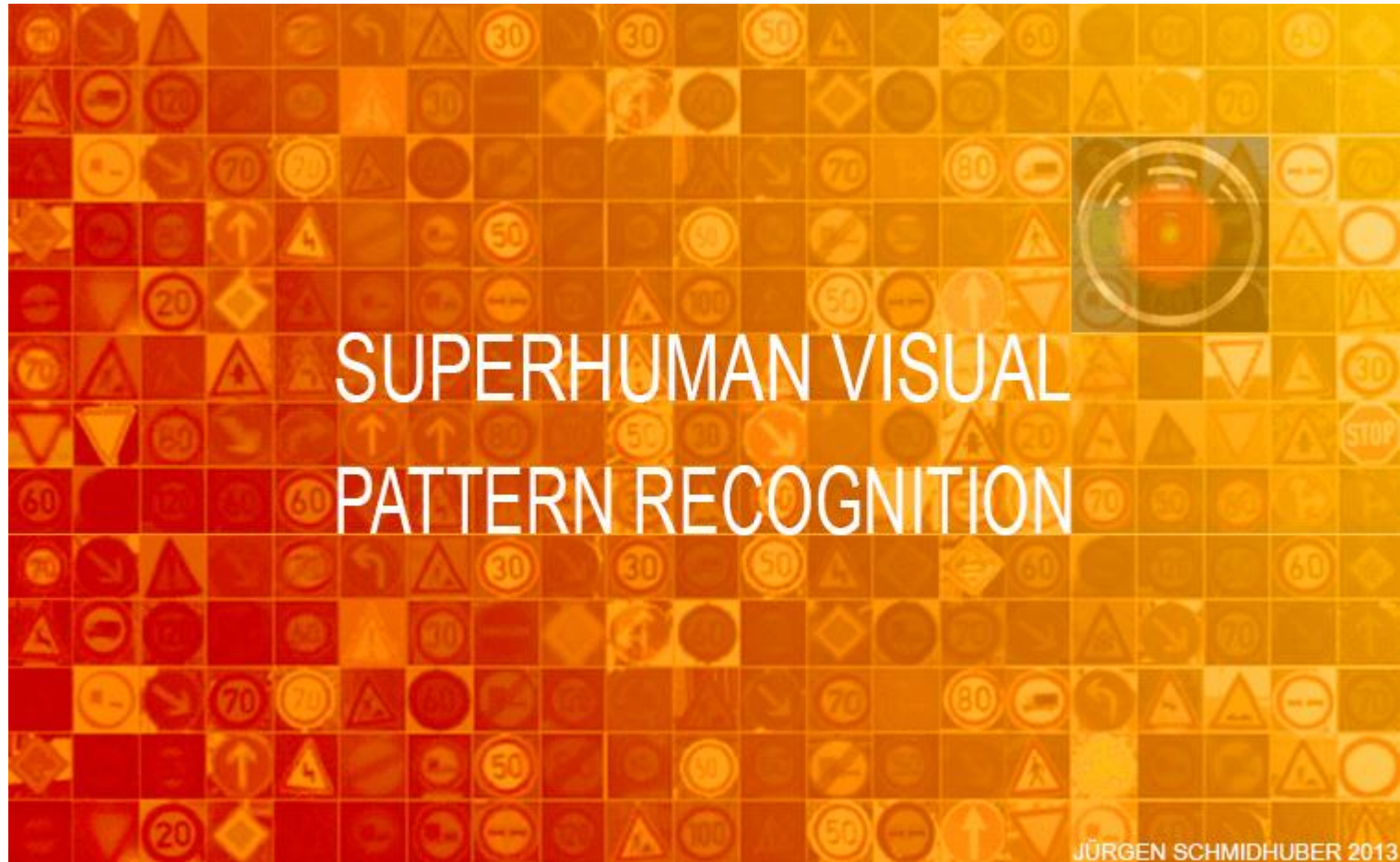
Joint Cascade Face Detection and Alignment,

Dong Chen, Shaoqing Ren, YichenWei, Xudong Cao, and Jian Sun, ECCV'2014

Распознавание изображений:

Convolution networks,
Deep learning,
Sparse coding,
Image Retrieval

Convolution networks, Deep learning, Image Retrieval



2011: First Superhuman Visual Pattern Recognition

twice better than humans

three times better than the closest artificial competitor

six times better than the best non-neural method

[Jürgen Schmidhuber](#)

<http://people.idsia.ch/~juergen/deeplearning.html>

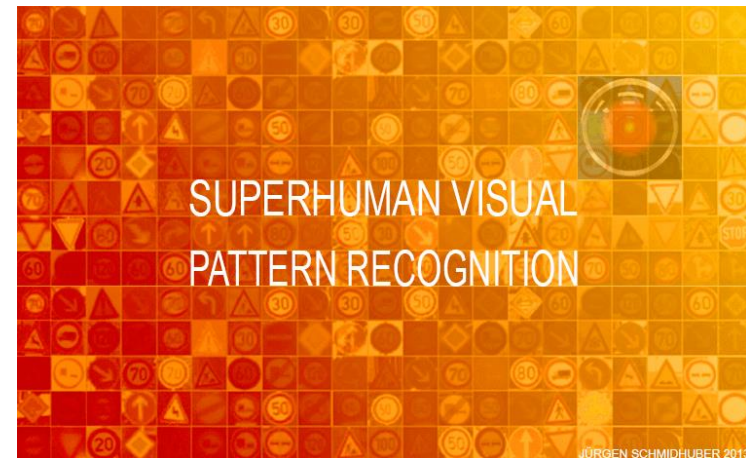
Convolution networks, Deep learning, Image Retrieval

The list of won competitions in Computer Vision:

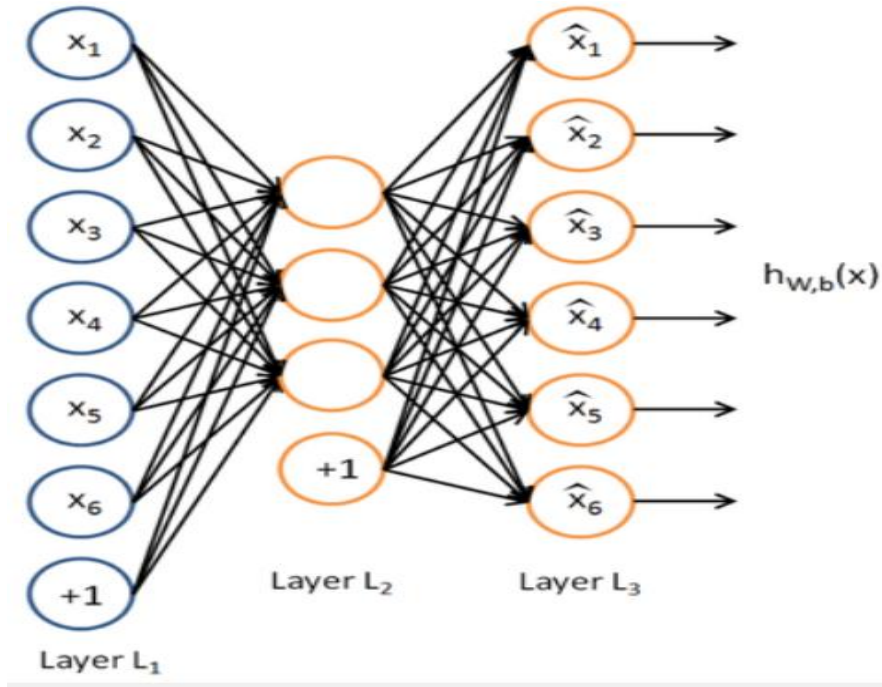
9. MICCAI 2013 Grand Challenge on Mitosis Detection
8. ICPR 2012 Contest on Mitosis Detection in Breast Cancer Histological Images
7. ISBI 2012 Brain Image Segmentation Challenge (with superhuman pixel error rate)
6. IJCNN 2011 Traffic Sign Recognition Competition (only our method achieved superhuman results)
5. ICDAR 2011 offline Chinese Handwriting Competition
4. Online German Traffic Sign Recognition Contest
3. ICDAR 2009 Arabic Connected Handwriting Competition
2. ICDAR 2009 Handwritten Farsi/Arabic Character Recognition Competition
1. ICDAR 2009 French Connected Handwriting Competition. Compare the overview page on handwriting recognition.

Records in important Machine Learning (ML) benchmarks:

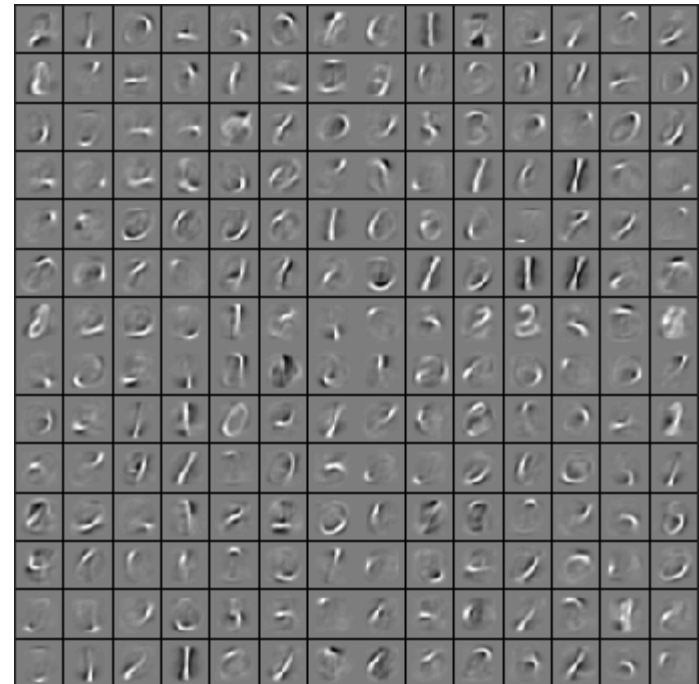
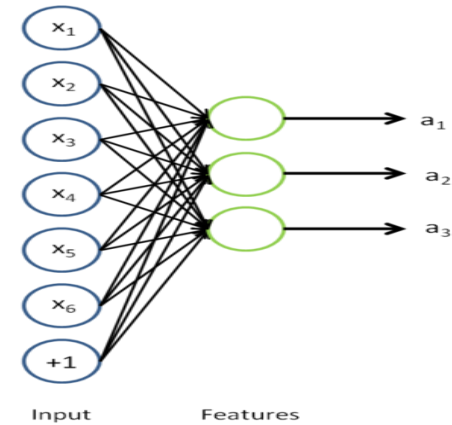
- D. Chinese characters from the ICDAR 2013 competition (3755 classes)
- C. The MNIST Handwritten Digits Benchmark (perhaps the most famous ML benchmark; we achieved the 1st human-competitive result in 2011)
- B. The CIFAR Image Classification Benchmark
- A. The NORB Stereo Vision Object Recognition Benchmark



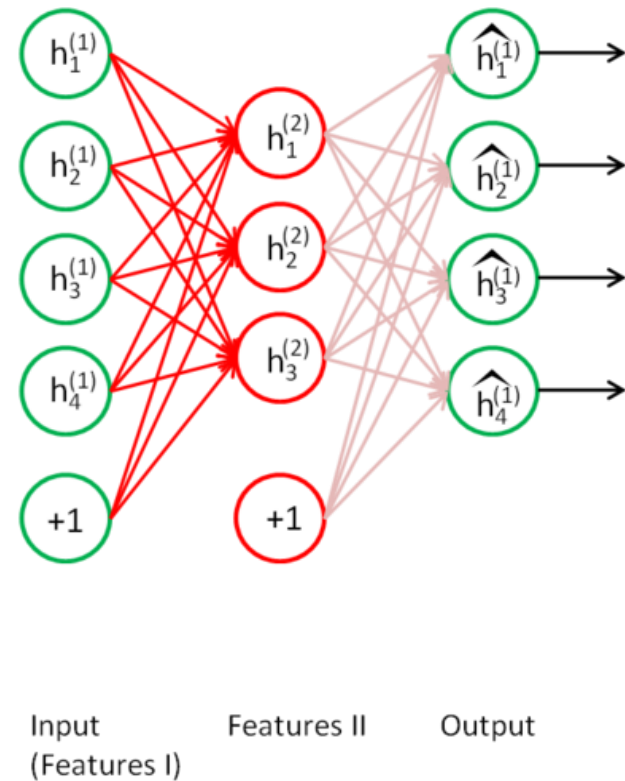
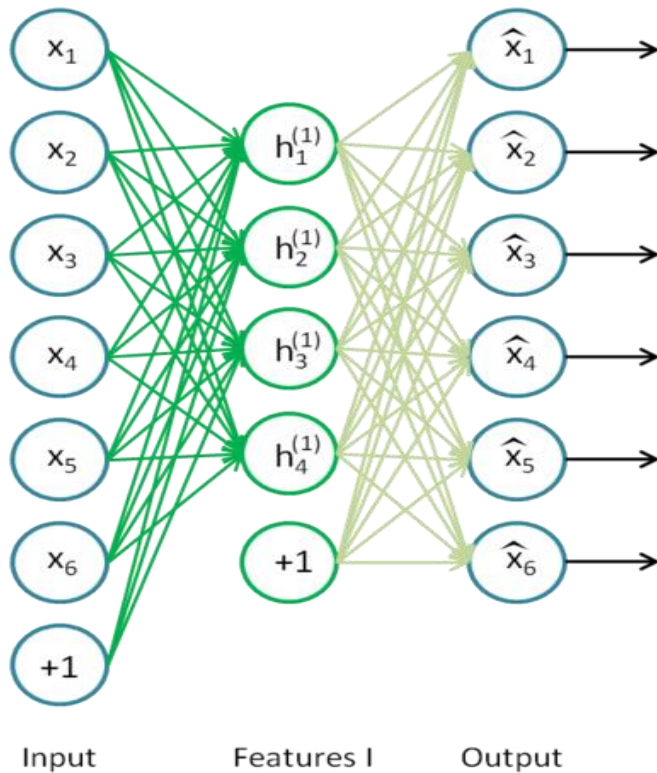
Convolution networks, Deep learning, Image Retrieval



Автоэнкодер

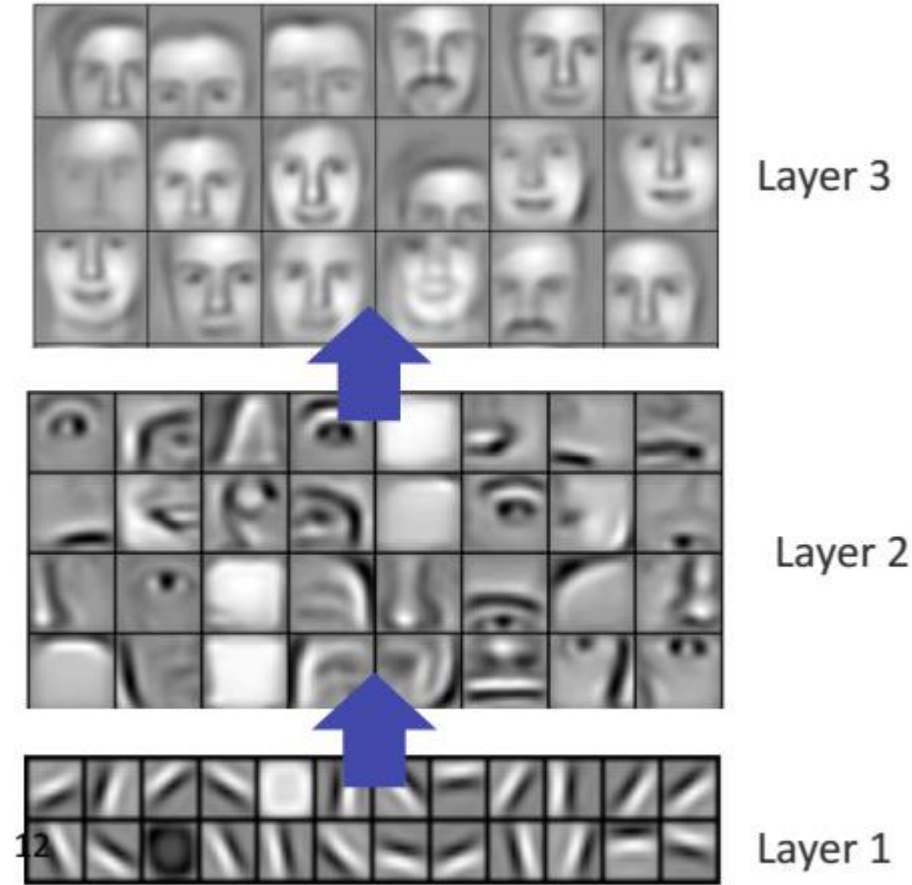
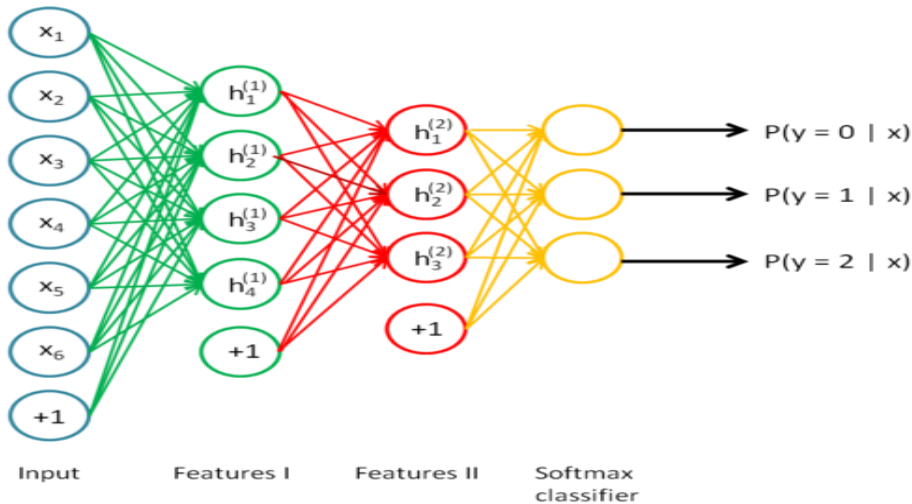
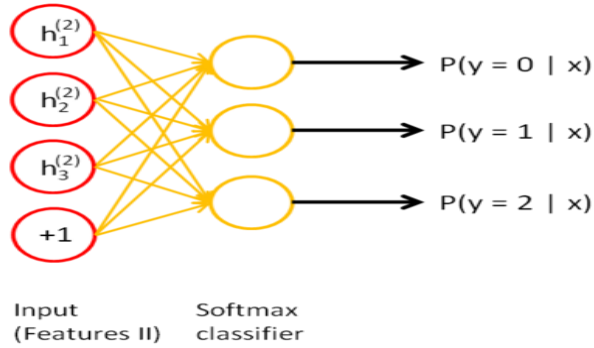


Convolution networks, Deep learning, Image Retrieval



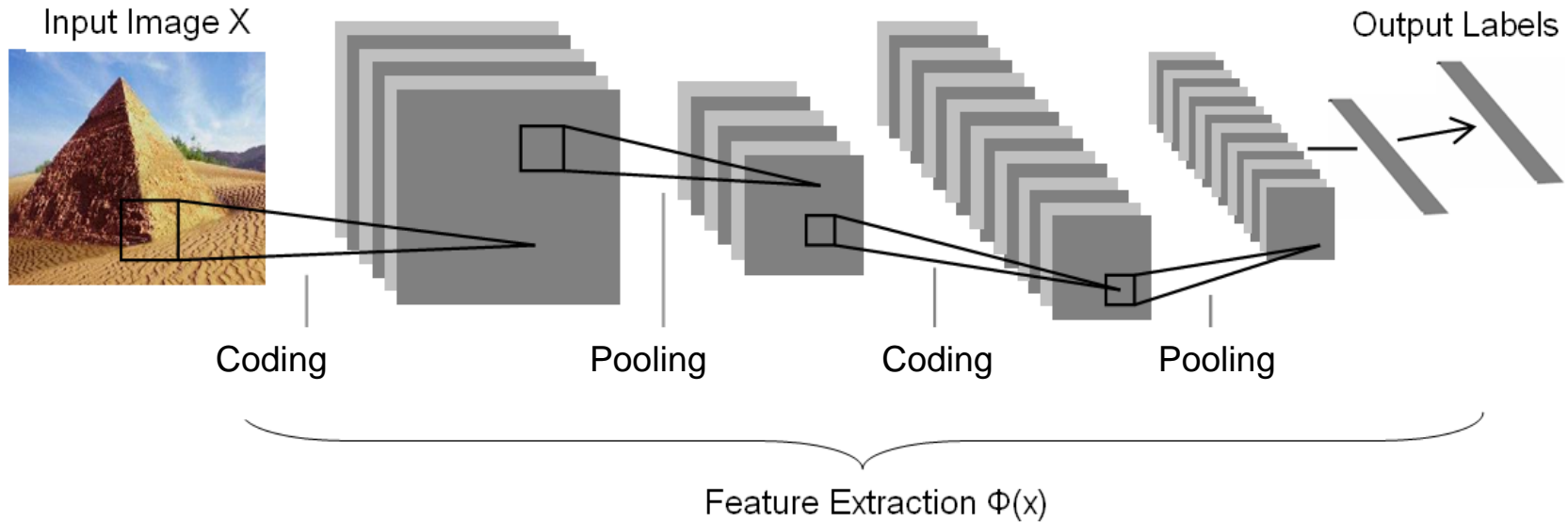
Deep architecture: обучение последовательности автоэнкодеров

Convolution networks, Deep learning, Image Retrieval



Deep architecture: обучение многослойной сети автоэнкодеров

Convolution networks, Deep learning, Image Retrieval



Архитектура “Coding + Pooling” = “Кодирование + Объединение”
(e.g., convolutional neural net, HMAX, BoW, ...)

- Coding: nonlinear mapping data into another feature space
- Better coding methods: **sparse coding**, RBMs, auto-encoders

Convolution networks, Deep learning, Image Retrieval

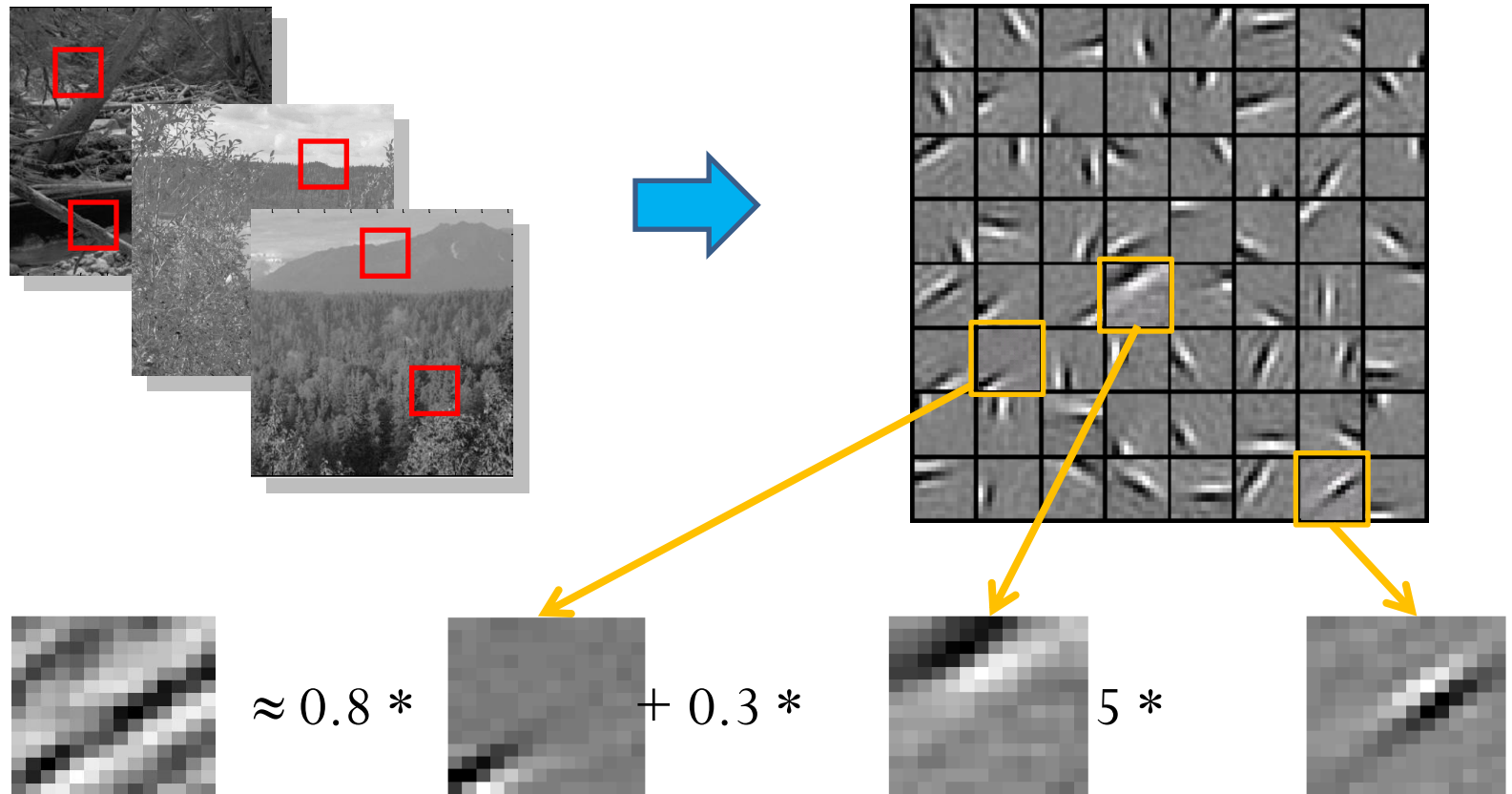
Sparse coding (Olshausen & Field, 1996). Originally developed to explain early visual processing in the brain (edge detection).

Training: given a set of random patches x , learning a dictionary of bases $[\Phi_1, \Phi_2, \dots]$

Coding: for data vector x , solve LASSO to find the sparse coefficient vector a

$$\min_{a, \phi} \sum_{i=1}^m \left\| x_i - \sum_{j=1}^k a_{i,j} \phi_j \right\|^2 + \lambda \sum_{i=1}^m \sum_{j=1}^k |a_{i,j}|$$

Convolution networks, Deep learning, Image Retrieval



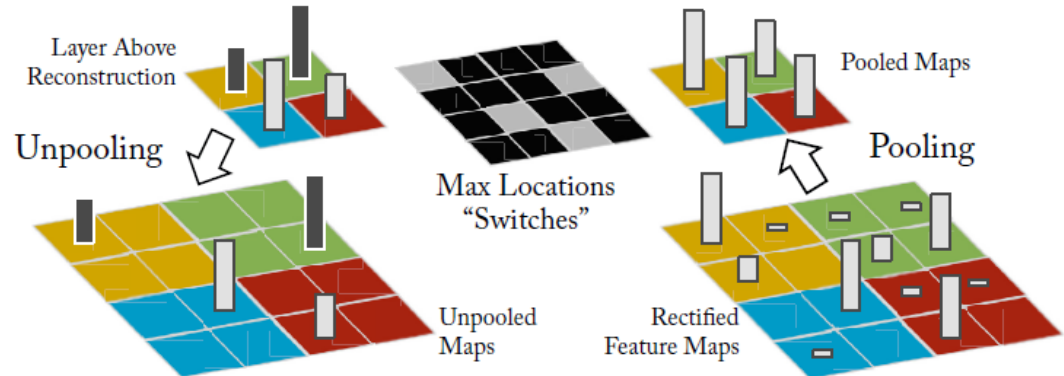
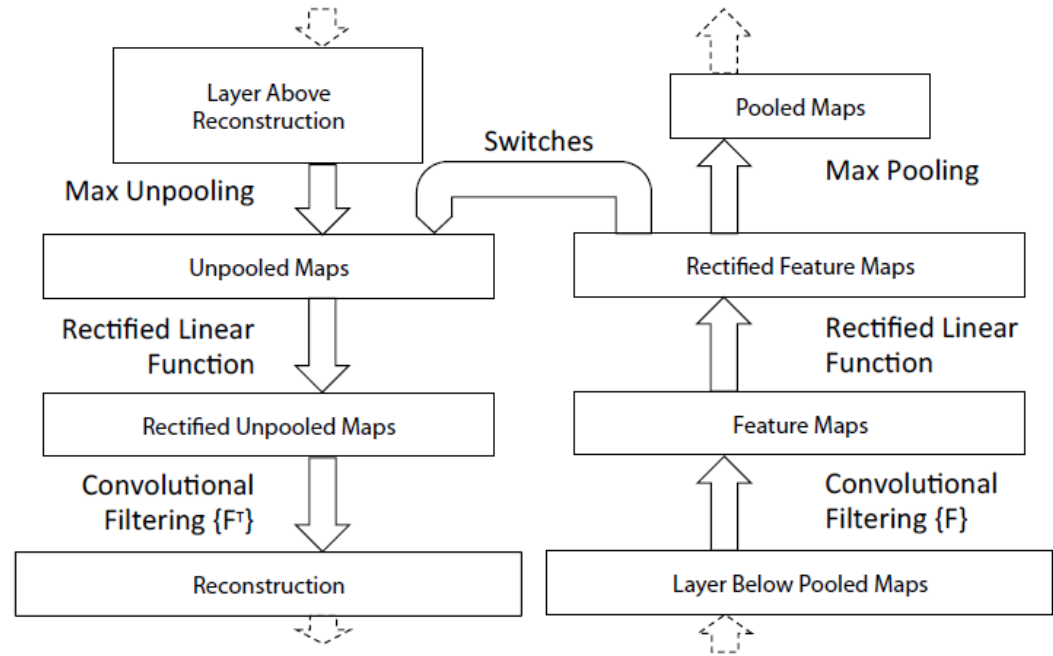
$[a_1, \dots, a_{64}] = [0, 0, \dots, 0, \mathbf{0.8}, 0, \dots, 0, \mathbf{0.3}, 0, \dots, 0, \mathbf{0.5}, 0]$
(feature representation)

Convolution networks, Deep learning, Image Retrieval

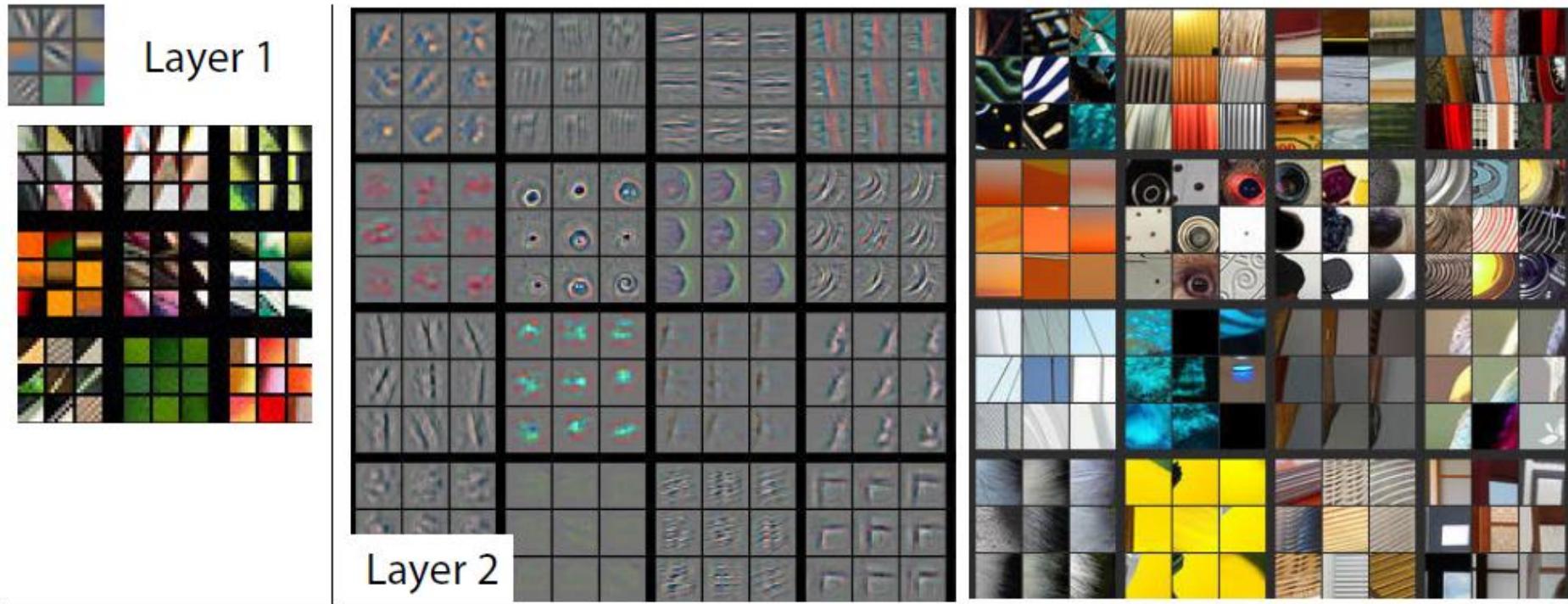
Novel visualization technique that gives insight into the function of intermediate feature layers and the operation of the classifier.

A deconvnet layer (left) attached to a convnet layer (right).

The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath.



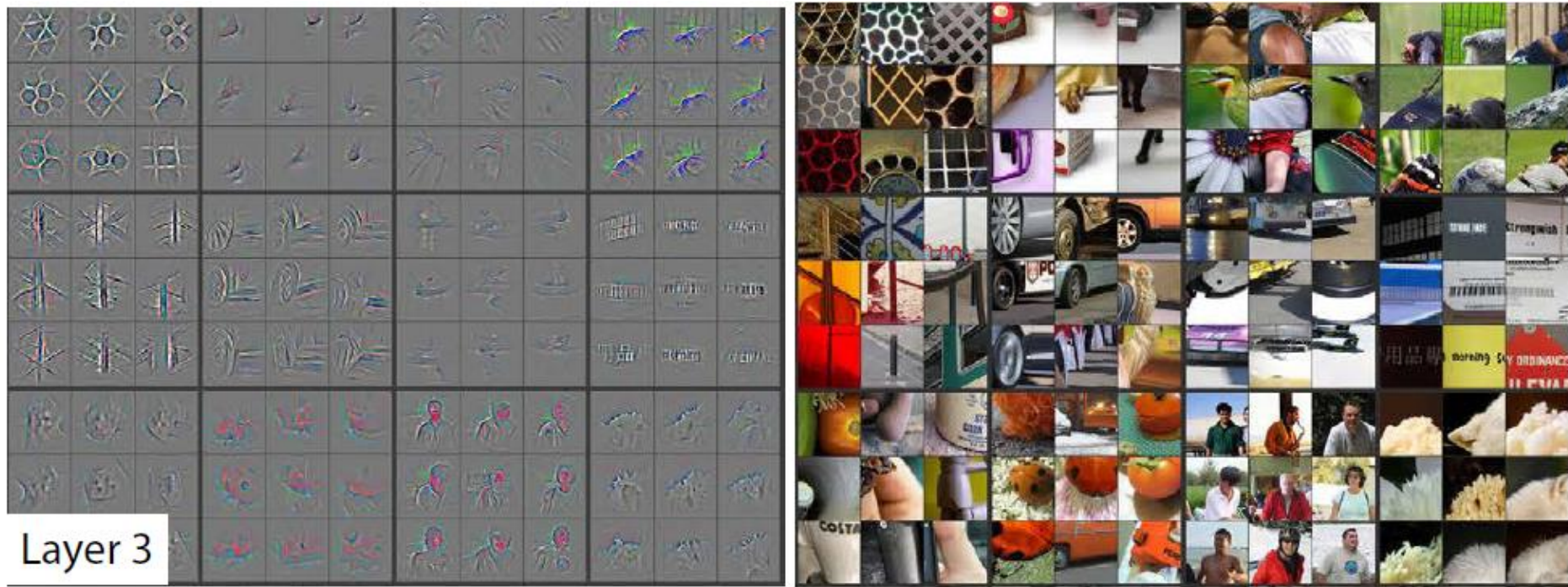
Convolution networks, Deep learning, Image Retrieval



Visualization of features in a fully trained model. For layers 2-5 we show the top 9 activations in a random subset of feature maps across the validation data, projected down to pixel space using our deconvolutional network approach. Our reconstructions are *not* samples from the model: they are reconstructed patterns from the validation set that cause high activations in a given feature map. For each feature map we also show the corresponding image patches.

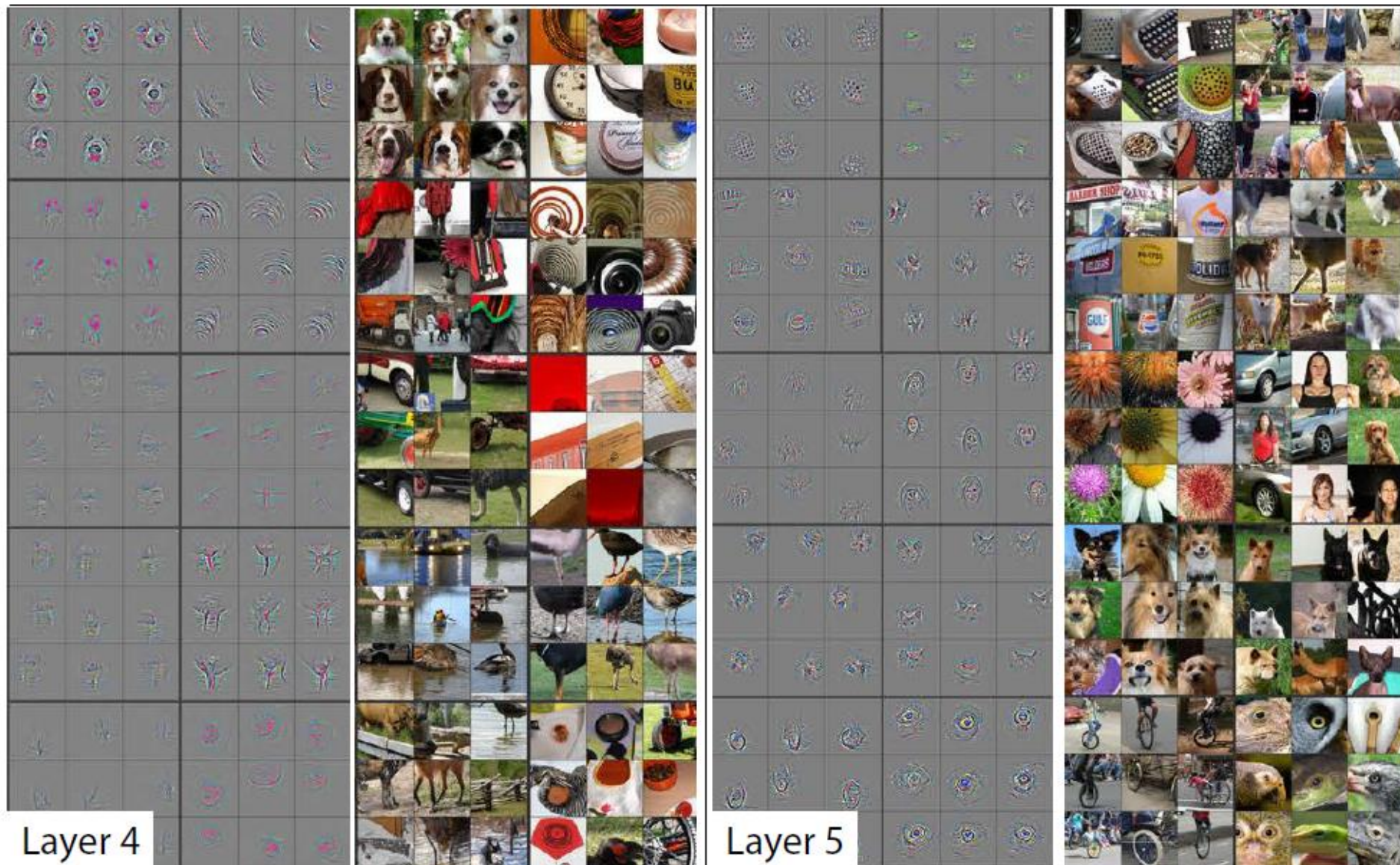
Visualizing and Understanding Convolutional Networks,
Matthew Zeiler and Rob Fergus, ECCV'14

Convolution networks, Deep learning, Image Retrieval



Visualizing and Understanding Convolutional Networks,
Matthew Zeiler and Rob Fergus, ECCV'14

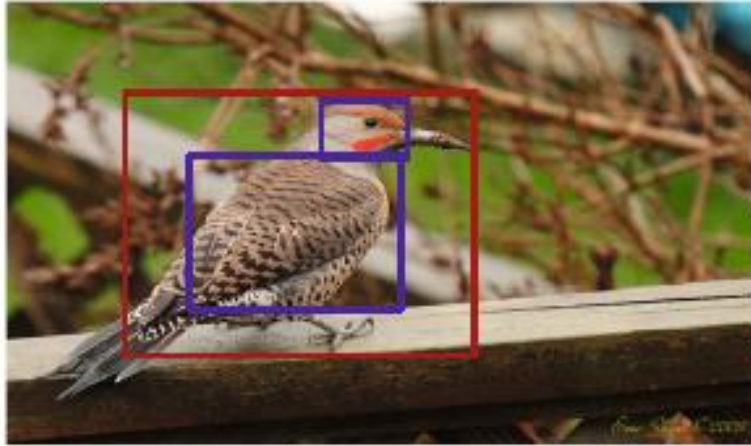
Convolution networks, Deep learning, Image Retrieval



Visualizing and Understanding Convolutional Networks,
Matthew Zeiler and Rob Fergus, ECCV'14

Convolution networks, Deep learning, Image Retrieval

Object detection and part localizations



Pose-normalized representation

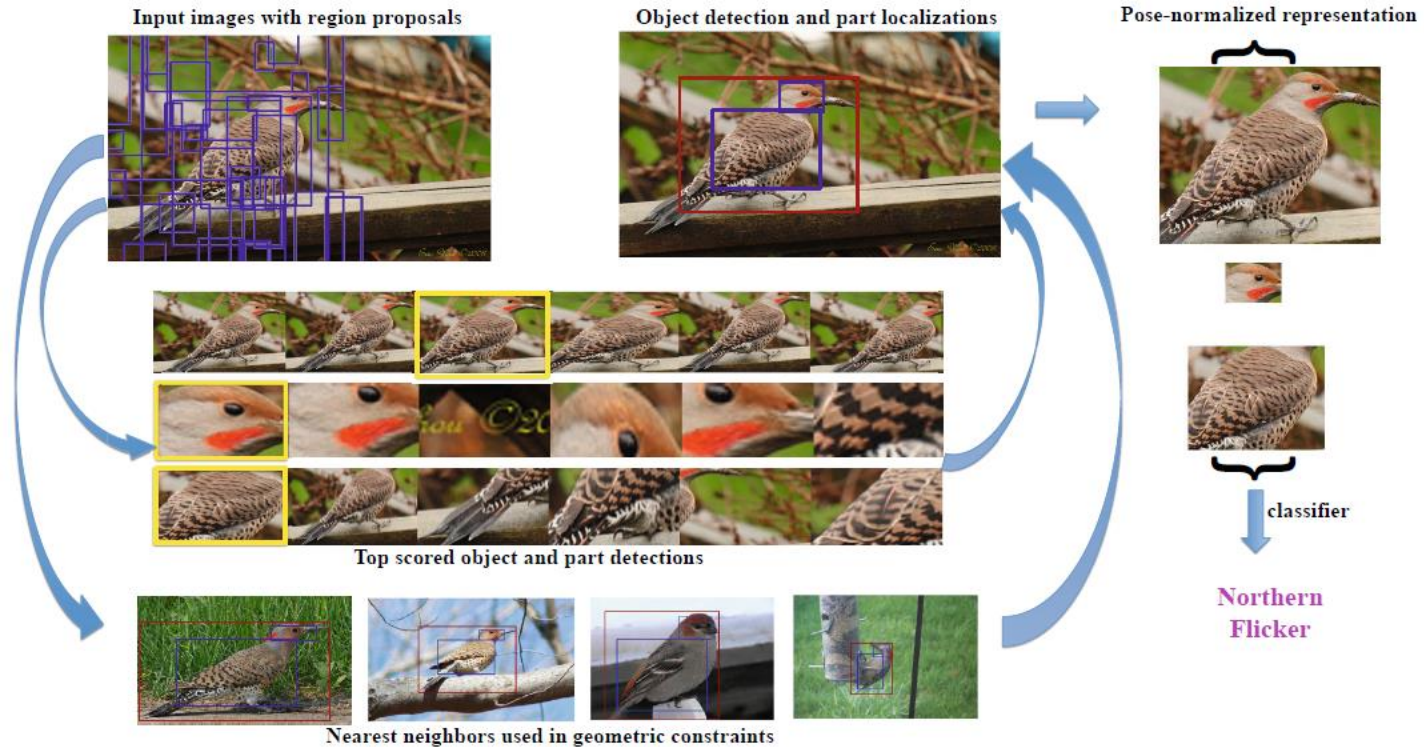


Специфика задачи детальной классификации объектов

The problem of **visual fine-grained categorization** can be extremely challenging due to the subtle differences in the appearance of certain parts across related categories. In contrast to basic-level recognition, fine-grained categorization aims to distinguish between different breeds or species or product models, and often requires **distinctions that must be conditioned on the object pose for reliable identification**. **Facial recognition is the classic case of fine-grained recognition**, and it is noteworthy that the best facial recognition methods jointly discover facial landmarks and extract features from those locations.

Part-based R-CNNs for Fine-Grained Category Detection,
Ning Zhang, Jeff Donahue, Ross Girshick, Trevor Darrell, ECCV'14

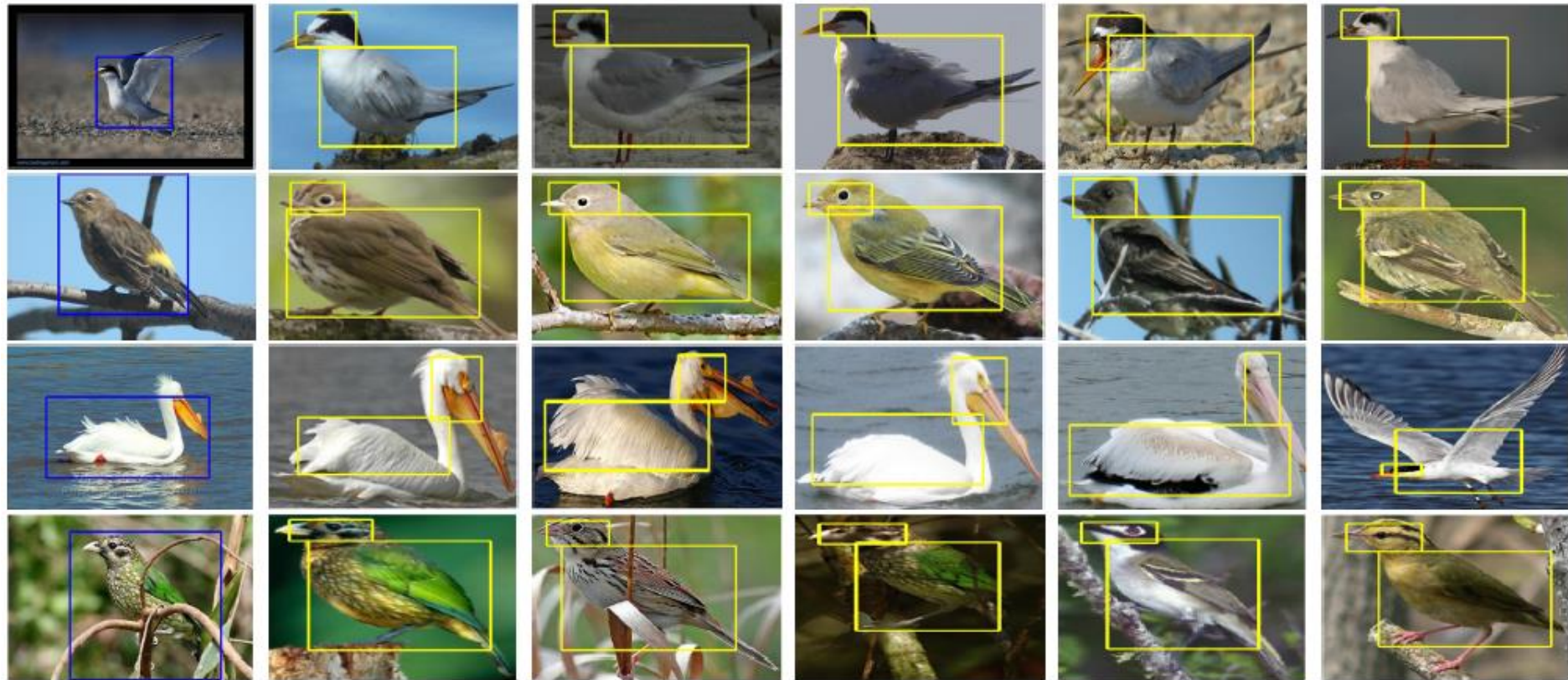
Convolution networks, Deep learning, Image Retrieval



Overview of our part localization: Starting from bottom-up region proposals (top-left), we train both object and part detectors based on deep convolutional features. During test time, all the windows are scored by all detectors (middle), and we apply non-parametric geometric constraints (bottom) to rescore the windows and choose the best object and part detections (top-right). The final step is to extract features on the localized semantic parts for fine-grained recognition for a pose-normalized representation and then train a classifier for the final categorization.

Part-based R-CNNs for Fine-Grained Category Detection,
Ning Zhang, Jeff Donahue, Ross Girshick, Trevor Darrell, ECCV'14

Convolution networks, Deep learning, Image Retrieval



In each row, the first column is the test image with an R-CNN bounding box detection, and the rest are the top-five nearest neighbors in the training set, indexed using pool5 features and cosine distance metric.

Part-based R-CNNs for Fine-Grained Category Detection,
Ning Zhang, Jeff Donahue, Ross Girshick, Trevor Darrell, ECCV'14

Convolution networks, Deep learning, Image Retrieval

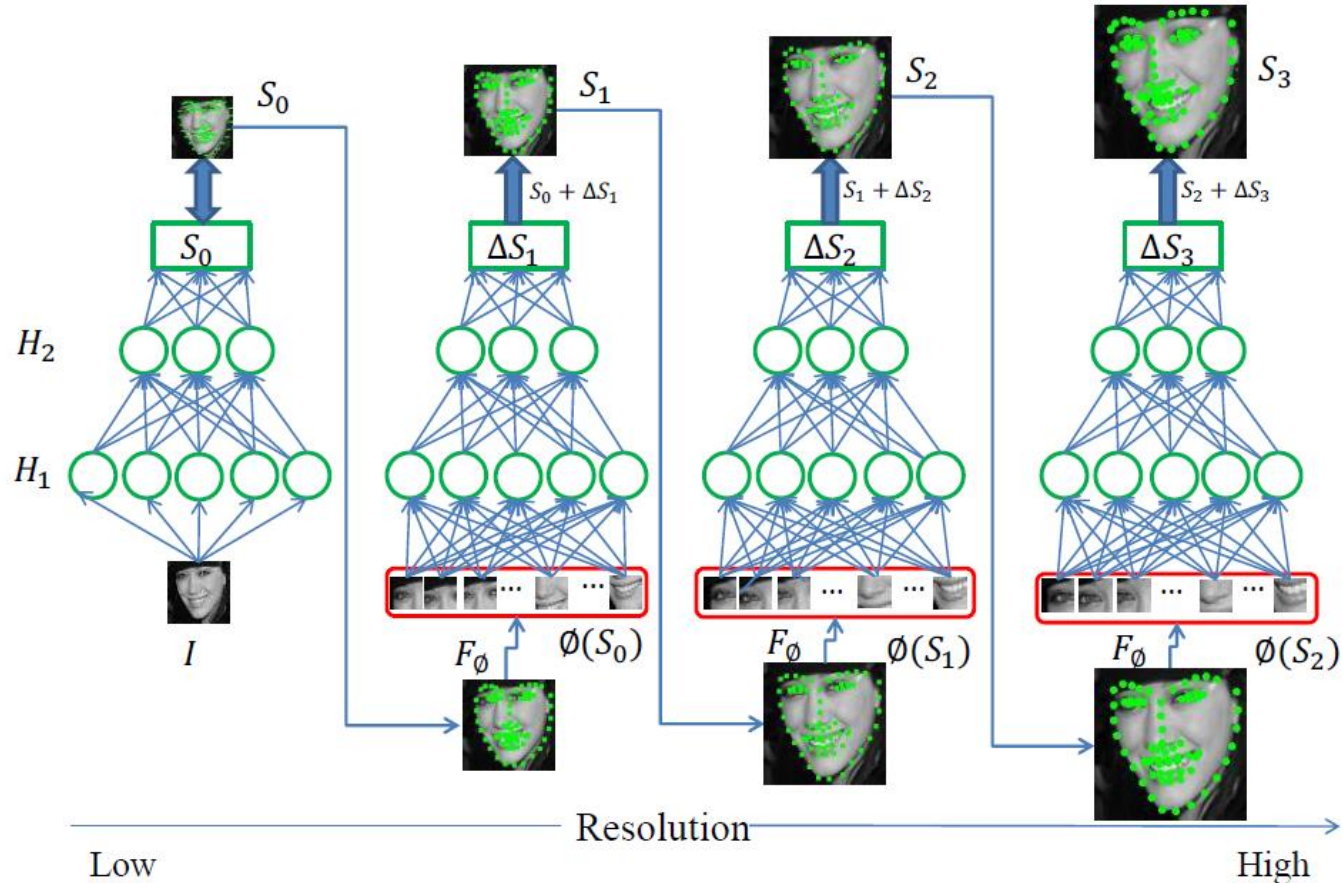


Fig. 1. Overview of our Coarse-to-Fine Auto-encoder Networks (CFAN) for real-time face alignment. H_1, H_2 are hidden layers. Through function F_Φ , the joint local features $\Phi(S_i)$ are extracted around facial landmarks of current shape S_i .

Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment,
Jie Zhang, Shiguang Shan, Meina Kan, Xilin Chen, ECCV'14

Convolution networks, Deep learning, Image Retrieval



Fig. 2. Facial landmark detection under the partial occlusion scenario (from Helen datasets [18]): Results of DCNN [26] (top row) and our CFAN (bottom row)



Fig. 3. Denition of 68 (top) and 49 (bottom) facial landmarks

Proposed CFAN attempts to design the general cascade regression framework in a coarse-to-fine architecture, with the regression in each stage modeled as a nonlinear deep network. Specifically, the CFAN framework consists of several successive *Stacked Auto-encoder Networks* (SANs). Each SAN attempts to characterize the nonlinear mappings from face image to face shape in different scales based on the shape predicted from the previous SAN.

Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment,
Jie Zhang, Shiguang Shan, Meina Kan, Xilin Chen, ECCV'14

Convolution networks, Deep learning, Image Retrieval

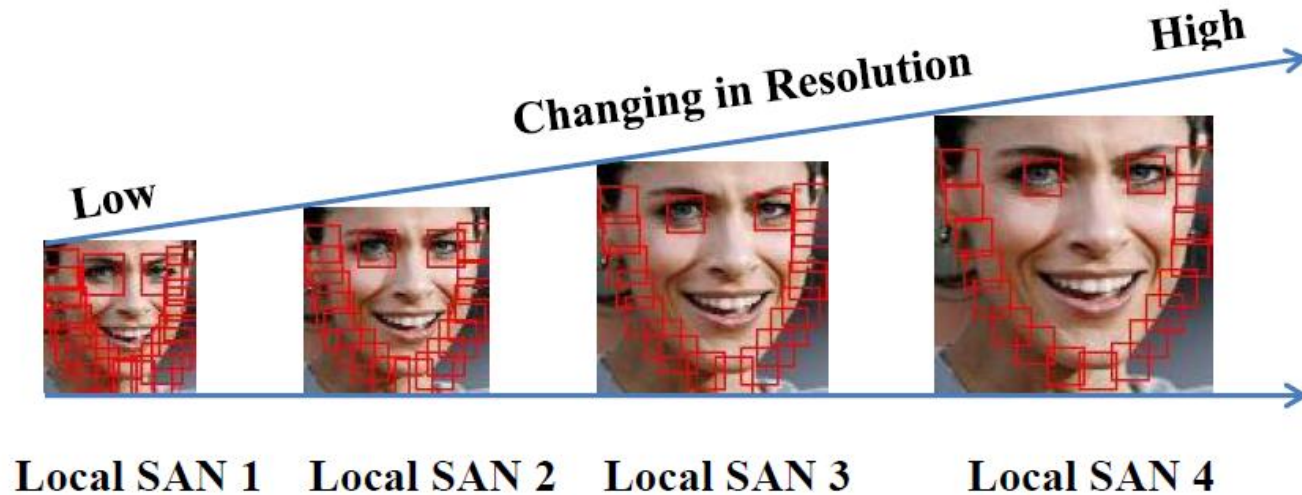


Fig. 4. Local patches extracted around the landmark points with different resolutions. For the sake of concise display, we choose two eye centers and 17 facial points on the face contour to describe the multi-resolution strategy used in each local SAN.

Convolution networks, Deep learning, Image Retrieval



Fig. 10. Example results from XM2VTS, LFPW and HELEN. The first five column samples contain diverse variations in pose, expression, beard, sunglasses and occlusion respectively. Some failure cases are shown in the last column.

Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment,
Jie Zhang, Shiguang Shan, Meina Kan, Xilin Chen, ECCV'14

Image Retrieval

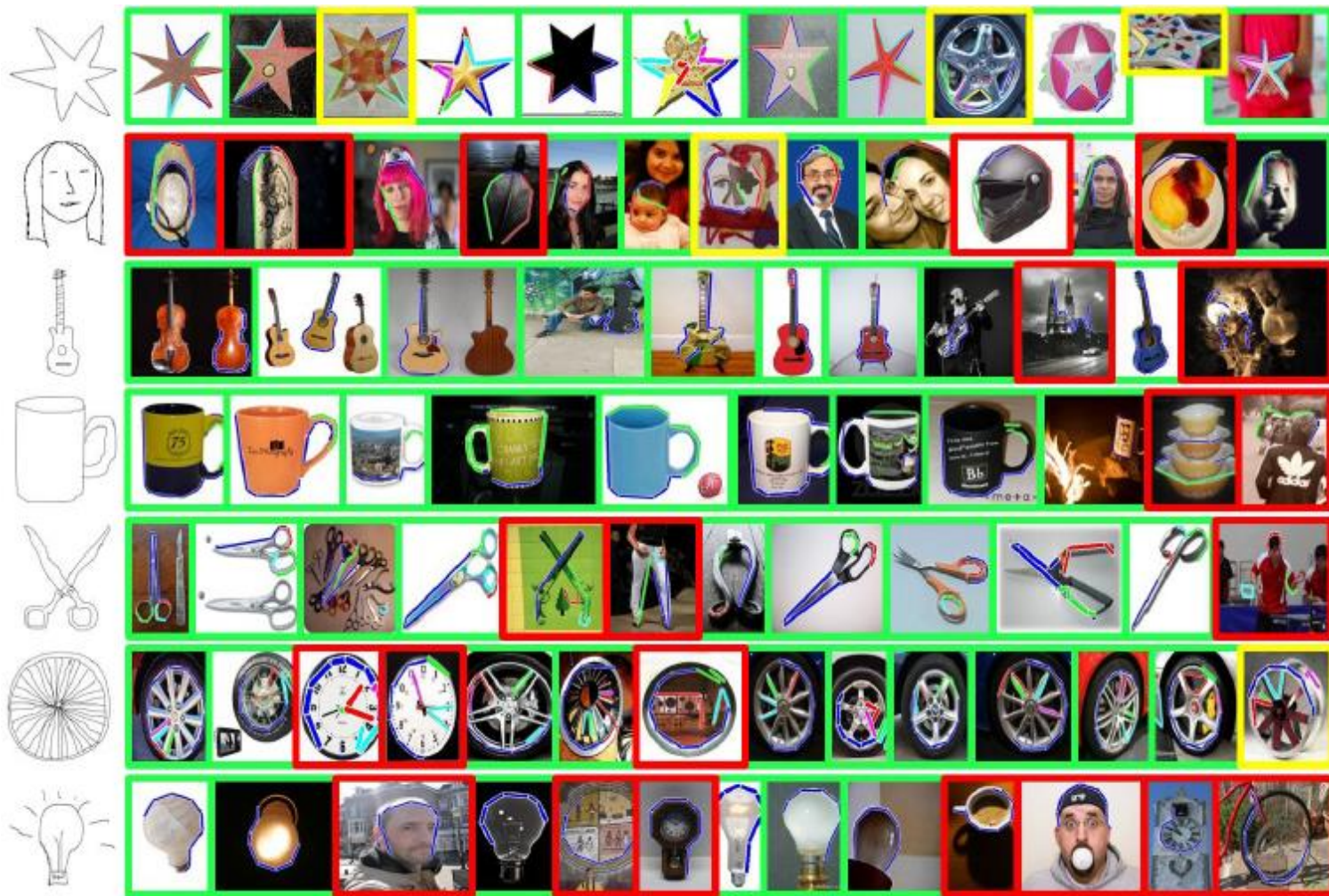
- Поиск по сходству в коллекциях изображений



Similarity-Invariant Sketch-Based Image Retrieval in Large Databases,
Sarthak Parui and Anurag Mittal, ECCV'14

Image Retrieval

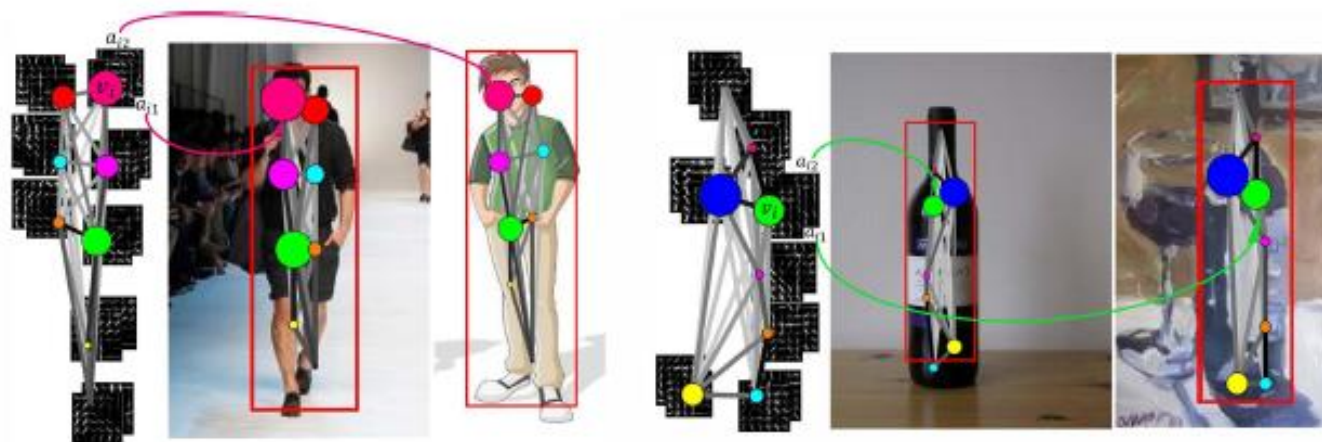
- Поиск по сходству в коллекциях изображений



Similarity-Invariant Sketch-Based Image Retrieval in Large Databases,
Sarthak Parui and Anurag Mittal, ECCV'14

Image Retrieval

- Поиск по сходству в коллекциях изображений

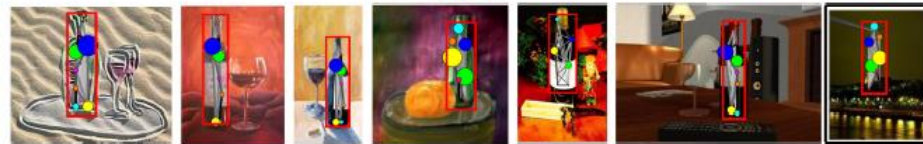


Person

Bike

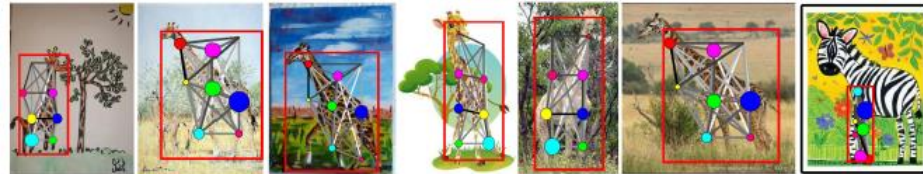


Bottle



Horse

Giraffe



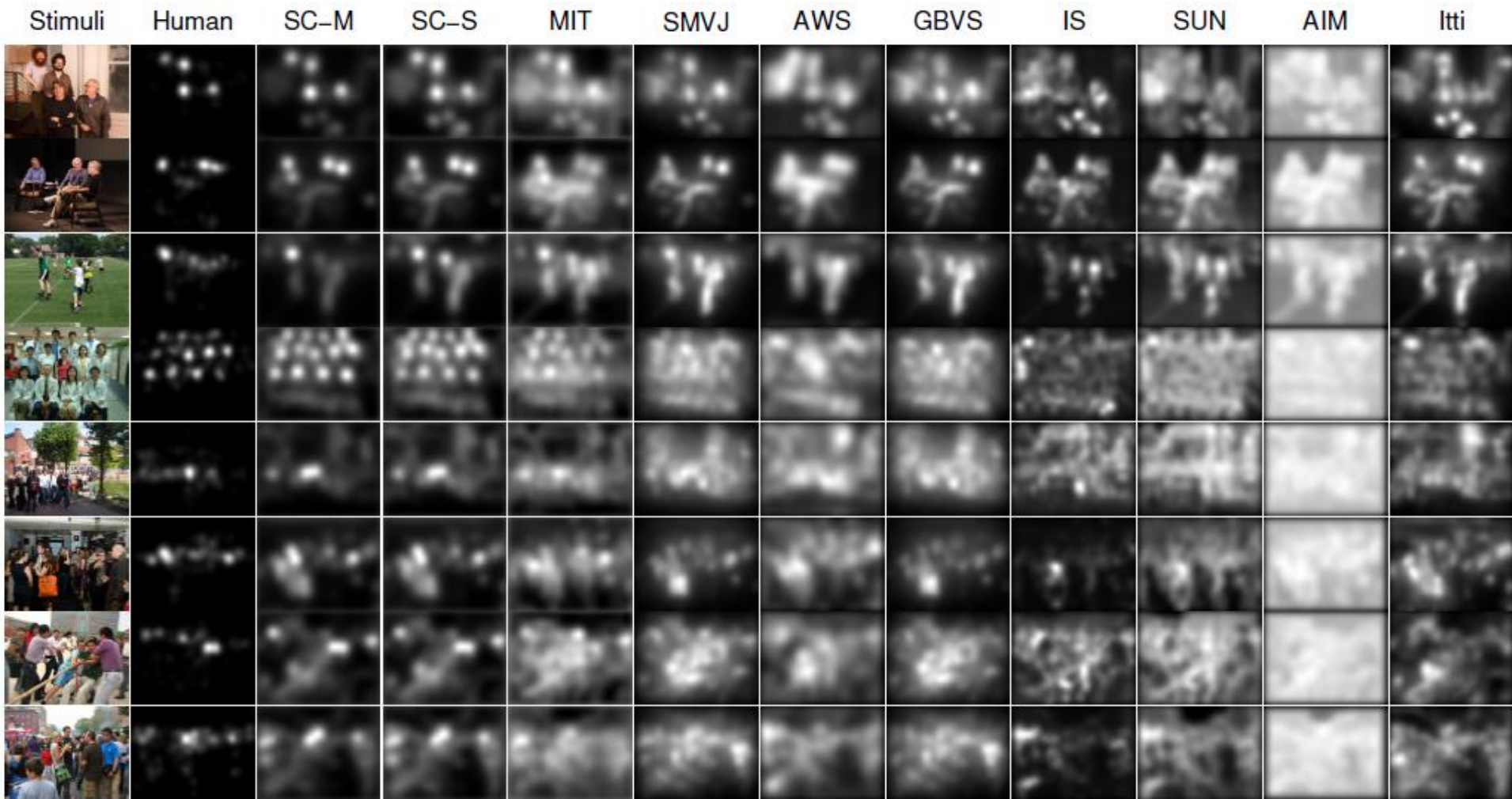
Learning Graphs to Model Visual Objects across Different Depictive Styles,
Qi Wu, Hongping Cai, and Peter Hall, ECCV'2014

Зрительное

внимание:

Saliency

Saliency



Saliency in Crowd,

Ming Jiang, Juan Xu, and Qi Zhao, ECCV'14

Saliency

Saliency is primarily driven in a bottom-up manner, depending on low level visual cues in the visual scene. In one of the first biologically plausible computational models for controlling visual attention, Koch and Ullman [31] followed Treisman and Gelade [46] and introduced the idea of a saliency map. Visual input is first decomposed into several maps encoding early visual features. Spatial competition in terms of hierarchical center-surround differences then determines their convergence to a unique map encoding saliency at each location. Most subsequent bottom-up saliency algorithms followed this model and compute the saliency of pixel constituents based on their local context (i.e., neighborhood) at multiple scales [27,22,10,25]. Alternatively, context was also considered globally, e.g., as a smoothed version of the amplitude [23] or the phase [20] spectrum of the image.

- 31. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry 4(4), 219–227 (1985)
- 46. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. *Cognitive Psychology* 12(1), 97–136 (1980)
- 27. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11), 1254–1259 (1998)
- 22. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: *Neural Information Processing Systems*, pp. 545–552 (2006)
- 10. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: *Neural Information Processing Systems*, pp. 155–162 (2005)
- 25. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. In: *Neural Information Processing Systems*, pp. 547–554 (2005)
- 23. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007)
- 20. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing* 19(1), 185–198 (2010)

A Closer Look at Context: From Coxels to the Contextual Emergence of Object Saliency

Rotem Mairon and Ohad Ben-Shahar, ECCV'14

Saliency

Unlike the models mentioned above, that mainly act spatially in order to reproduce human visual search strategies or predict visual fixations, other methods aim at detecting saliency at the higher level of *objects*. While the (local) visual context used by the first class of methods is reasonably intuitive, the forms of visual context employed by the latter (object-level) approaches typically remain unexplained. We argue that this somewhat obscure relationship often constrains the nature of visual objects they may capture in order to measure their saliency.

Definition. The *visual context* of a constituent is the set of visual units in the image that are used in the computational process that measures its saliency.

Saliency

Contrast-Based Saliency: In the first group are approaches that associate saliency with high contrast between local or regional structures. To measure this contrast, the computational mechanisms employ various center-surround structures. Some approaches define the surround component independent of visual content, e.g., as the local neighborhood of a pixel [24,48,1,32] or larger regular blocks [33]. In other approaches, the surrounding context depends on a grouping process which typically results in a superpixel representation of the image [29,11]. Apart from reducing computational costs, superpixels are preferable due to their capacity to preserve locally coherent structures (unlike pixels or predefined blocks). To a certain extent, these structures facilitate meaningful central constituents when measuring contrast and therefore are more suitable for saliency assignment.

24. Hu, Y., Xie, X., Ma, W.-Y., Chia, L.-T., Rajan, D.: Salient region detection using weighted feature maps based on the human visual attention model. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3332, pp. 993–1000. Springer, Heidelberg (2004)

48. Wang, L., Xue, J., Zheng, N., Hua, G.: Automatic salient object extraction with contextual cue. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 105–112 (2011)

1. Achanta, R., Estrada, F.J., Wils, P., Sussstrunk, S.: Salient region detection and segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008)

32. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2), 353–367 (2011)

33. Ma, Y., Zhang, H.: Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of the Eleventh ACM international conference on Multimedia, pp. 374–381 (2003)

29. Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S.: Automatic salient object segmentation based on context and shape prior. In: British Machine Vision Conference (2011)

11. Chang, K., Liu, T., Chen, H., Lai, S.: Fusing generic objectness and visual saliency for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 914–921 (2011)

A Closer Look at Context: From Coxels to the Contextual Emergence of Object Saliency

Rotem Mairon and Ohad Ben-Shahar, ECCV'14

Saliency

Rarity-Based Saliency: The second group of approaches consider saliency as distinctness or rarity. Intuitively, these may signal the importance of a visual constituent compared with the redundancy of recurring visual information. Often in this approach the context is a global representation of the entire visual input. For example, such a representation may be the image mean color vector that is used as reference to measure the saliency at all other pixels [2,4]. Alternative representation has considered a smoothed version of the phase spectrum [28] in order to suppress non-salient components in the original spectrum and thus highlight salient locations after transforming back to the spatial domain. In a somewhat related way, image patches that are highly dissimilar to their k-nearest neighbors were considered salient as this indicates their dissimilarity to all other patches [19,11]. Recently, this measure of dissimilarity has been shown oblivious to patch statistics, leading to a new measure based on the distance of each patch to the average patch along the principal components of the patch distribution [35].

- 2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604 (2009)
- 4. Achanta, R., Susstrunk, S.: Saliency detection for content-aware image resizing. In: Proceedings of the IEEE International Conference on Image Processing, pp. 1005–1008 (2009)
- 11. Chang, K., Liu, T., Chen, H., Lai, S.: Fusing generic objectness and visual saliency for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 914–921 (2011)
- 35. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013)

A Closer Look at Context: From Coxels to the Contextual Emergence of Object Saliency

Rotem Mairon and Ohad Ben-Shahar, ECCV'14

Saliency

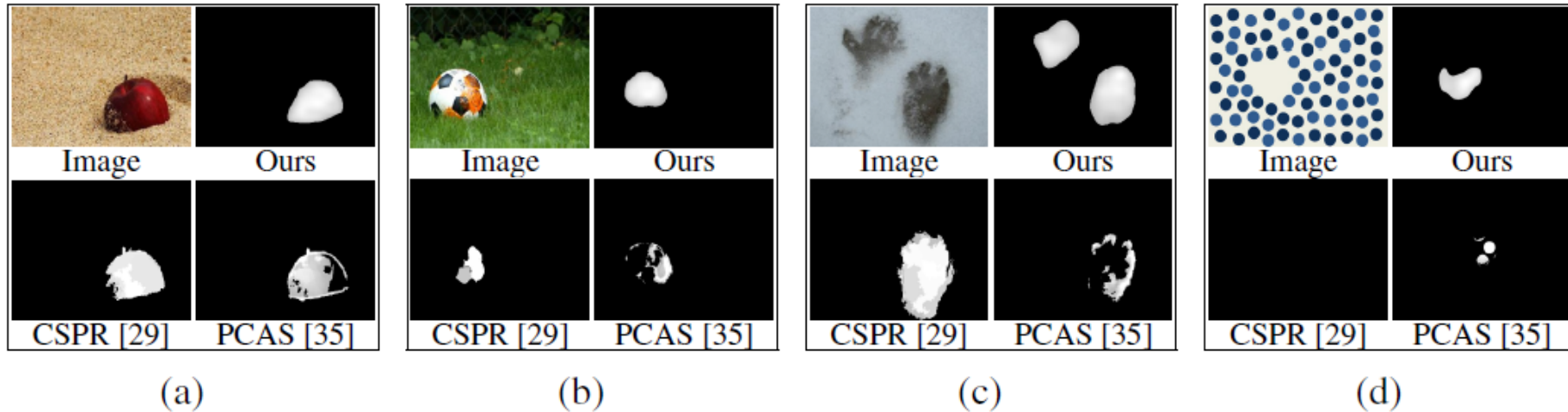
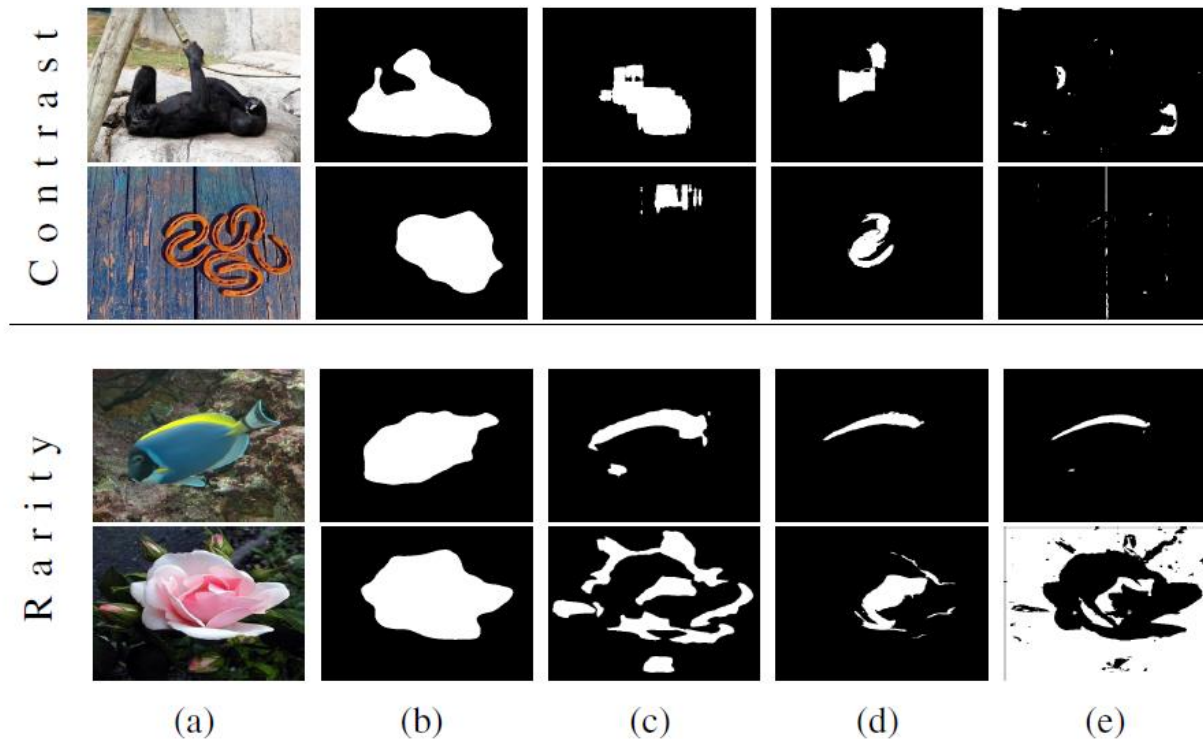


Fig. 1. Salient objects in visual stimuli can have different flavors. As is typical in virtually all benchmark databases, salient objects can be uniform singletons (panel a). However, salient objects can be multi-part and heterogeneous (panel b), they can have some multiplicity (panel c), or they can even be completely abstract (like the "hole" in panel d). By their implied notion of visual context, most computational saliency models impose certain constraints on the types of objects they can handle, with practical success limited to the simpler cases. Here we show computed saliency map (thresholded at 80%) from two state-of-the-art algorithms (CSPR [29] and PCAS [35]) and our own method. By modeling context instead of the objects we significantly reduce the constraints on the nature of objects that may be detected as salient, as is illustrated by the better assignment of saliency in all these cases.

A Closer Look at Context: From Coxels to the Contextual Emergence of Object Saliency

Rotem Mairon and Ohad Ben-Shahar, ECCV'14

Saliency



Binarized saliency maps demonstrate the challenges in capturing whole salient objects by contrast (top) and rarity (bottom) based approaches. The two leftmost columns in each category show example images and our maps. **Contrast:** Saliency maps in columns c and d are generated as part of saliency computation algorithms. In column c computation is based on rectangular structures of varying size and aspect ratio [32] whereas in column d neighboring superpixels were used to estimate contrast [29]. The constraints are even more restrictive when only local considerations are involved [1] as shown in column e. **Rarity:** The challenge remains when relying on rarity aspects of saliency, as demonstrated by the maps in columns c-e [19,35,14]. When the object consists of multiple parts, only those with rare appearance are detected. The bottom map in panel e demonstrates how a large object may render the appearance of its surrounding more rare and therefore more computationally salient.

A Closer Look at Context: From Coxels to the Contextual Emergence of Object Saliency

Rotem Mairon and Ohad Ben-Shahar, ECCV'14

Saliency

What are the characteristics of visual context which allow to consider the visual information it embeds (be it an object or not) as salient?

To answer this question, we suggest to model visual context based on the several characteristics of visual information. Given a particular representation of the units that compose it (pixels, superpixels, patches, etc...), we consider a single *context element*, or *coxel*, to be a region or a subset of the image with the following properties (see Fig. 4):

Smoothness: Nearby units that compose the coxel are expected to have similar visual appearance. The more distant the units, more leeway is allowed in their similarity.

Apathy to contiguity: A coxel may be either contiguous or not, i.e., it may constitute several distinct connected components in the image plane.

Enclosure: To qualify as a saliency coxel, the spatial layout of the context element should “enclose” (strictly or approximately) some visual information.

Saliency

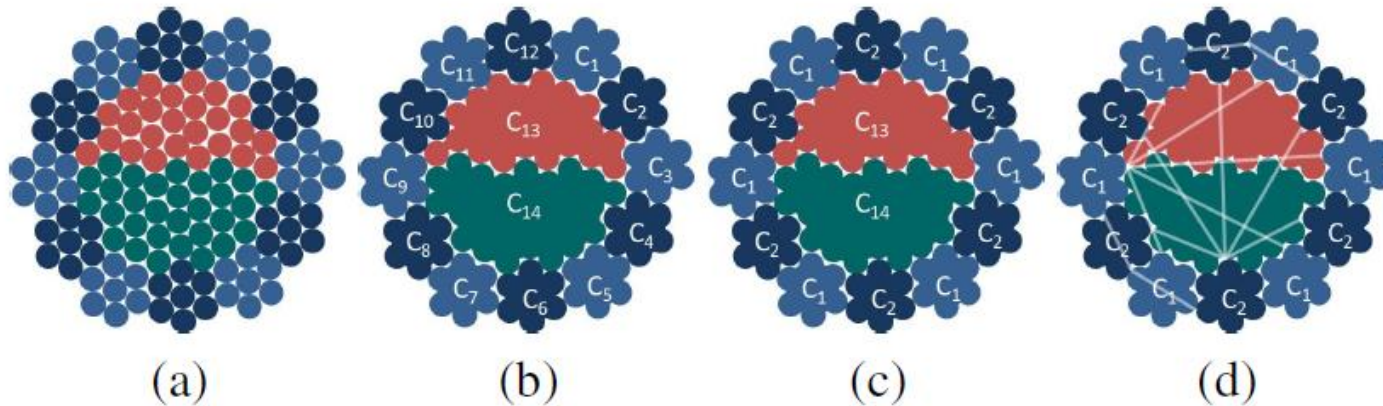
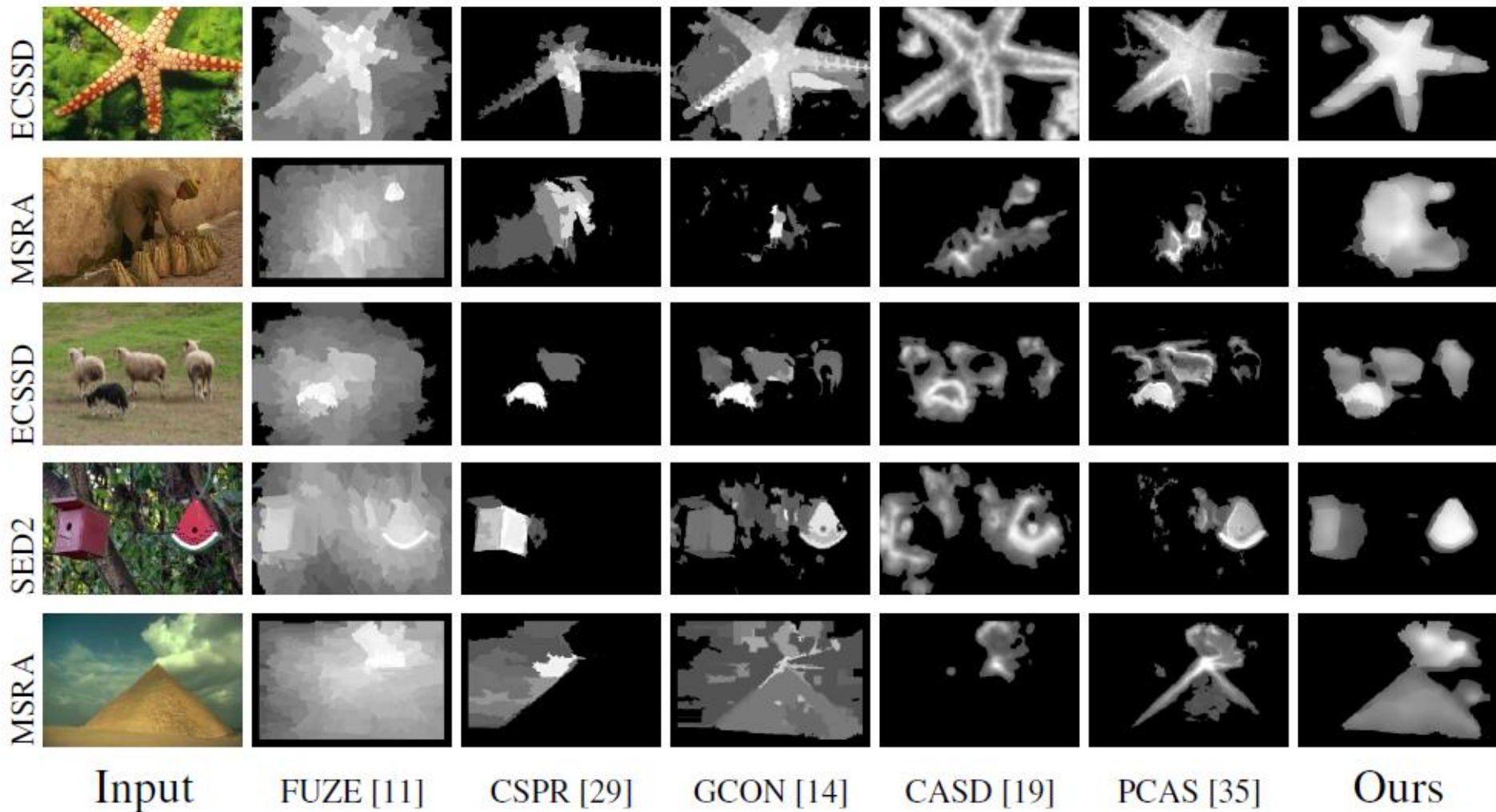


Fig. 5. Schematic depiction of the two phases of Algorithm 1. **(a)** Initial coxels (SLIC superpixels [3]) with their color-coded appearance content. **(b)** Coxels with small contextual gaps (initially, those which are very proximate and similar) are merged to larger, uniquely labeled components. Note that at this time no saliency bridges occur as any edge between two patches from the same component traverses another patch from that component. **(c)** At a future merging step, the threshold on contextual gaps is large enough to allow distant coxels to merge (implied by similar labels). **(d)** At this point, saliency bridges cross image patches from other coxels, leading to accumulation of their saliency measure. To avoid clutter, only selected number of saliency bridges are shown.

Saliency



A Closer Look at Context: From Coxels to the Contextual Emergence of Object Saliency

Rotem Mairon and Ohad Ben-Shahar, ECCV'14

Saliency

RGBD Salient Object Benchmark

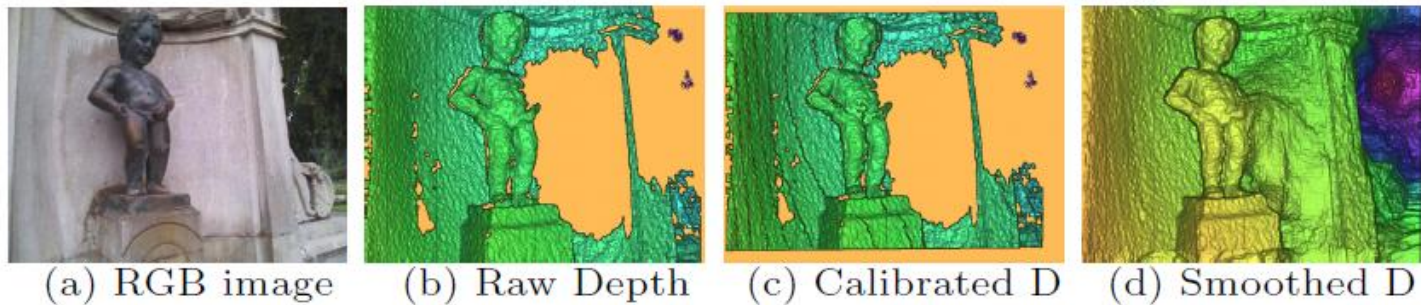
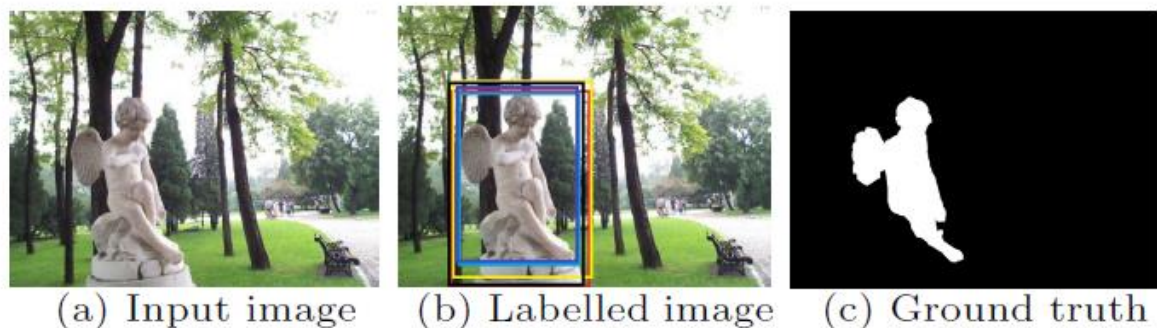


Fig. 1. Depth image calibration and filling

Fig. 2. A sample of image annotation. The image (b) is consistently labeled by five participants and included into our benchmark. (c) shows the final annotated salient object.



<http://sites.google.com/site/rgbdsaliency>

RGBD Salient Object Detection: A Benchmark and Algorithms,

Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji, ECCV'14

Saliency

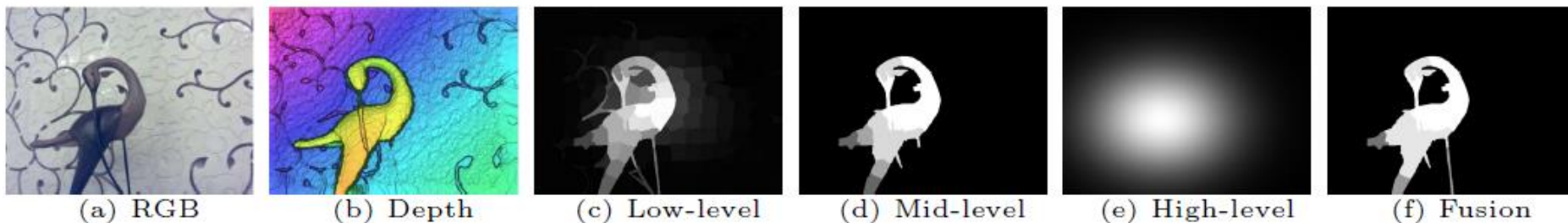
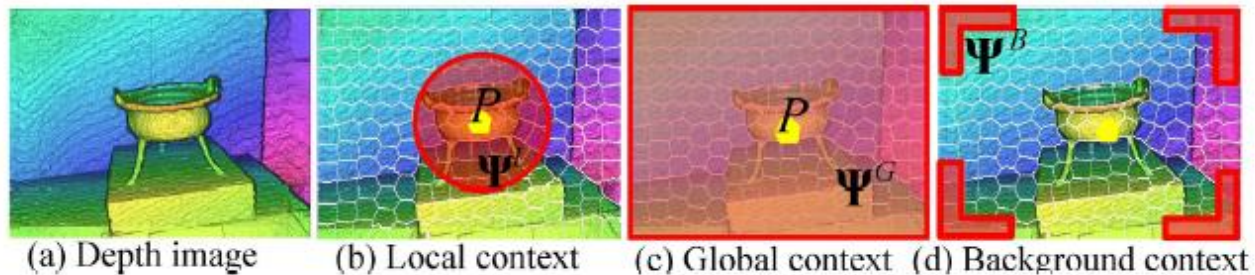
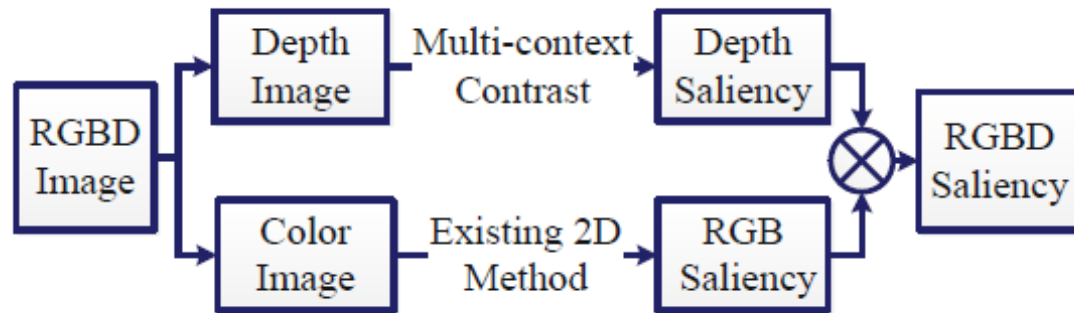
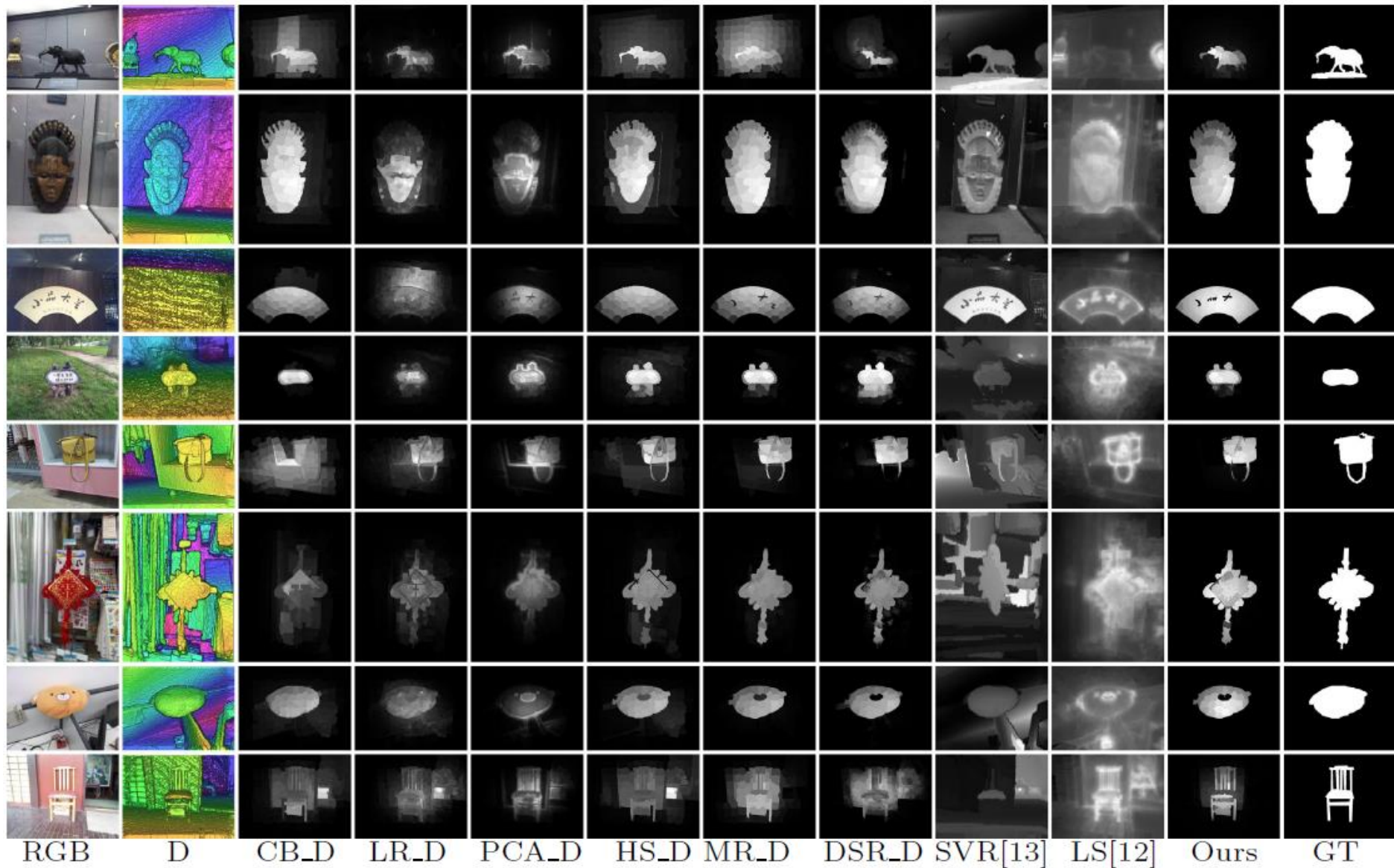


Fig. 8. Saliency maps produced by the key three stages in our approach

Saliency



RGBD Salient Object Detection: A Benchmark and Algorithms,
Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji, ECCV'14)

Сегментация:

Image, Video & 3D
Segmentation, MRF,
Energy-based,
Graphical Models,
Supercixels, 3D-Flow

Video Segmentation, 3D, Energy-based, Saliency

part – segmentation object – segmentation 3D – reconstruction

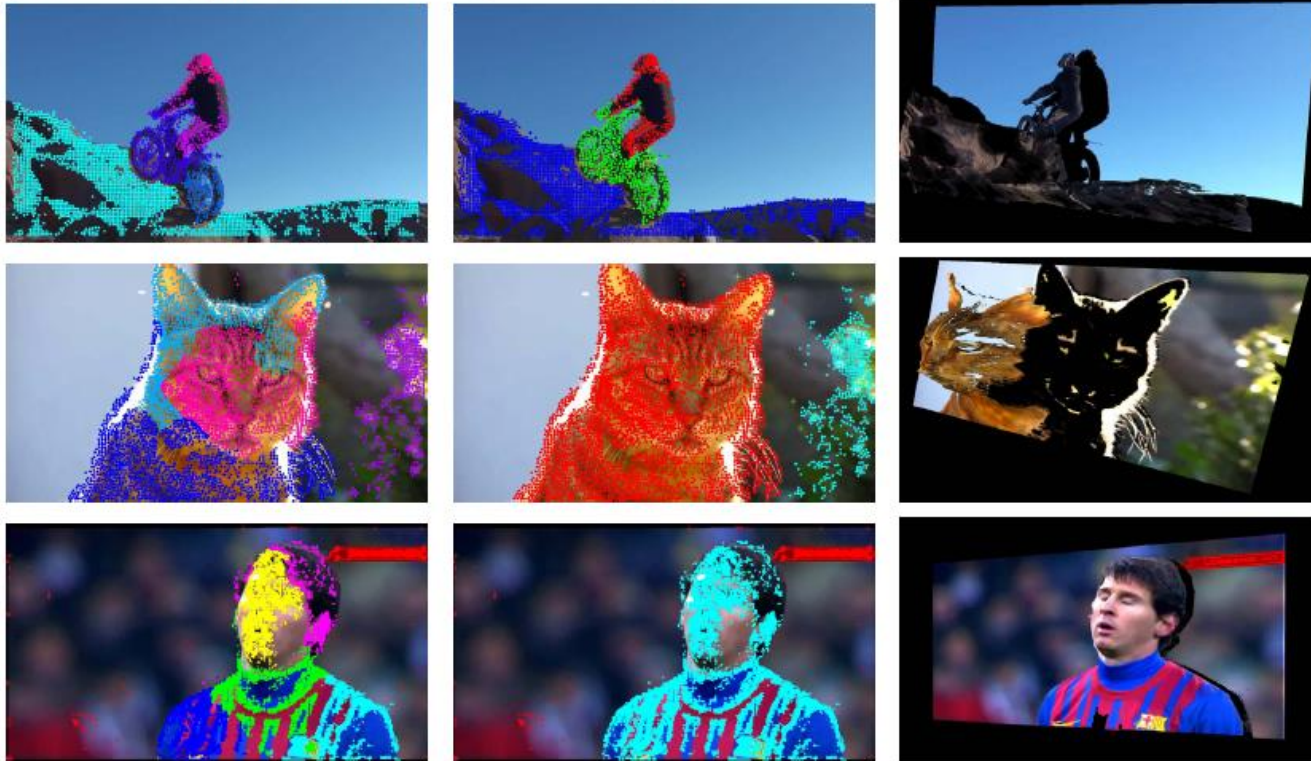


Fig. 1. Segmentation and 3D reconstruction results of two dynamic sequences of the *Youtube-Objects* Dataset [23] and a football sequence downloaded from YouTube. **Left:** segmentation into *parts* (rigid models). **Centre:** segmentation into *objects*. **Right:** densified 3D video pop-up from a novel viewpoint.

Video Segmentation, 3D, Energy-based, Saliency

$$\begin{aligned} C(\mathbf{x}) &= E_{data} + E_{edge_break} + E_{sparse} + E_{mdl} \\ &= \sum_{i \in \mathcal{T}} \sum_{m \in \mathbf{x}_i} U_i(m) + \sum_{i \in \mathcal{T}} \sum_{j \in N_i} d_{i,j} \Delta(j \notin N'_i) \\ &\quad + \sum_{m \neq n \in \mathcal{M}} \Delta(\exists i : I_i = m, n \in \mathbf{x}_i) + \text{MDL}(\mathbf{x}) \end{aligned}$$

Unary Costs (E_{data}) $U_i(m) = G_i(m) + P_i(m)$

Saliency Term. The work [28] provides a fully unsupervised method for object detection in an image I , using a novel saliency map S_I . While [28] made use of both the statistics taken from a large corpus of unlabelled images, and from the image itself, we only make use of the statistics of the single image (this measure is termed *within image saliency* in [28]). We compute saliency maps S_{I_f} for each frame f in the video sequence and define the saliency cost $P_i(m)$ of point i belonging to model m as the distance from the mean saliency of model m

$$P_i(m) = \lambda_s \sum_{f \leq F} (S_{I_f}(i) - \bar{S}_m)^2$$

where \bar{S}_m is the mean saliency of all tracks that currently belong to model m , $S_{I_f}(i)$ is the saliency score of point i in frame f and λ_s a weight on the importance of this term.

28. Siva, P., Russell, C., Xiang, T., Agapito, L.: Looking beyond the image: Unsupervised learning for object saliency and detection. CVPR'13

Video Pop-up: Monocular 3D Reconstruction of Dynamic Scenes,

Chris Russell, Rui Yu, and Lourdes Agapito, ECCV'14

Video Segmentation, 3D, Energy-based, Saliency



Fig. 4. Reconstruction results for a cat sequence of the *Youtube-Objects* Dataset [23]

Video Pop-up: Monocular 3D Reconstruction of Dynamic Scenes,
Chris Russell, Rui Yu, and Lourdes Agapito, ECCV'14

Video Segmentation, 3D, Energy-based, Saliency

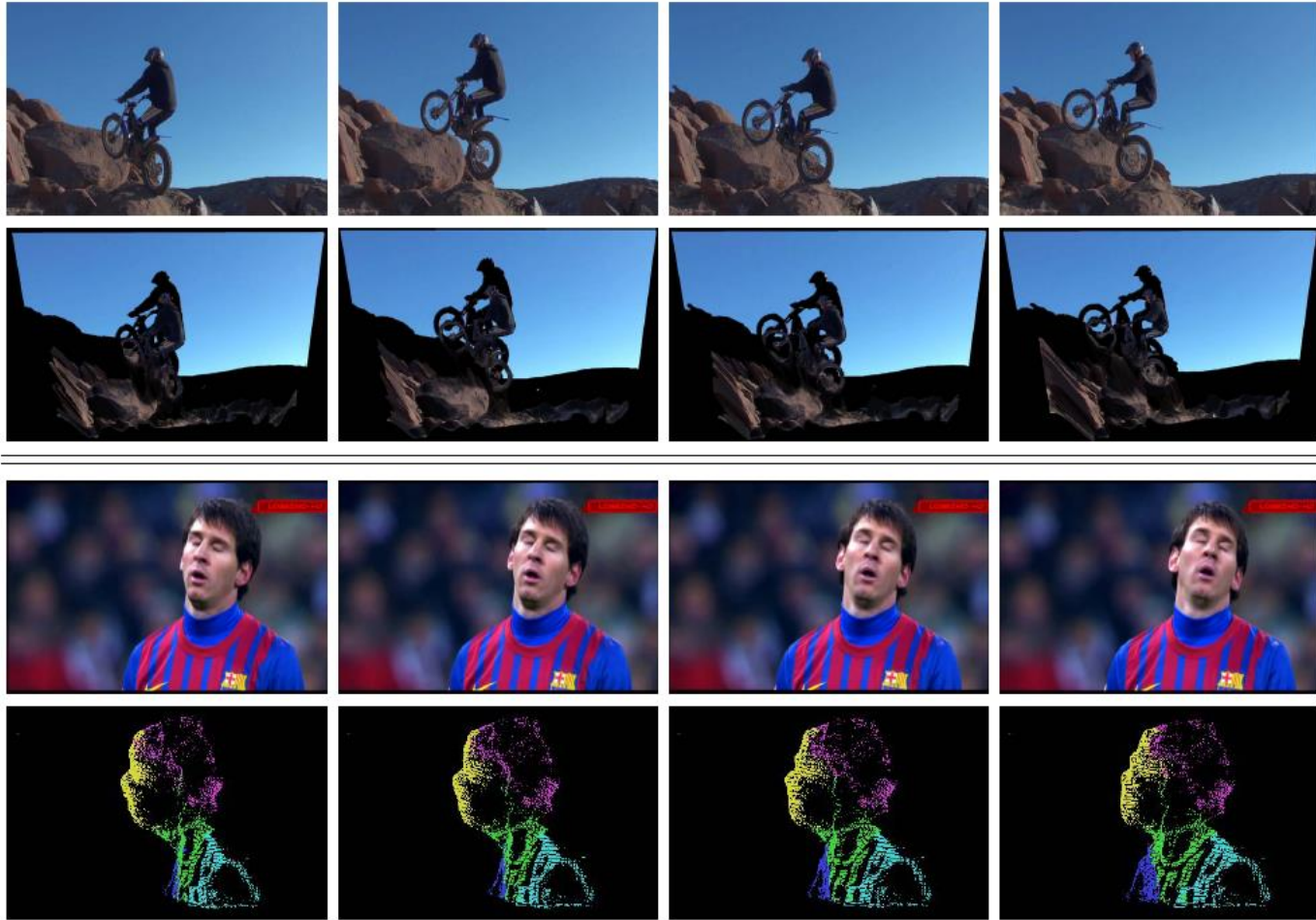


Fig. 5. **Top:** Reconstruction results for a motorbike sequence *Youtube-Objects* Dataset [23]. **Bottom:** Sparse reconstruction of football footage, showing both the assignment of tracks to parts and the quality of reconstruction before densification.

Video Segmentation, 3D, Energy-based, Saliency

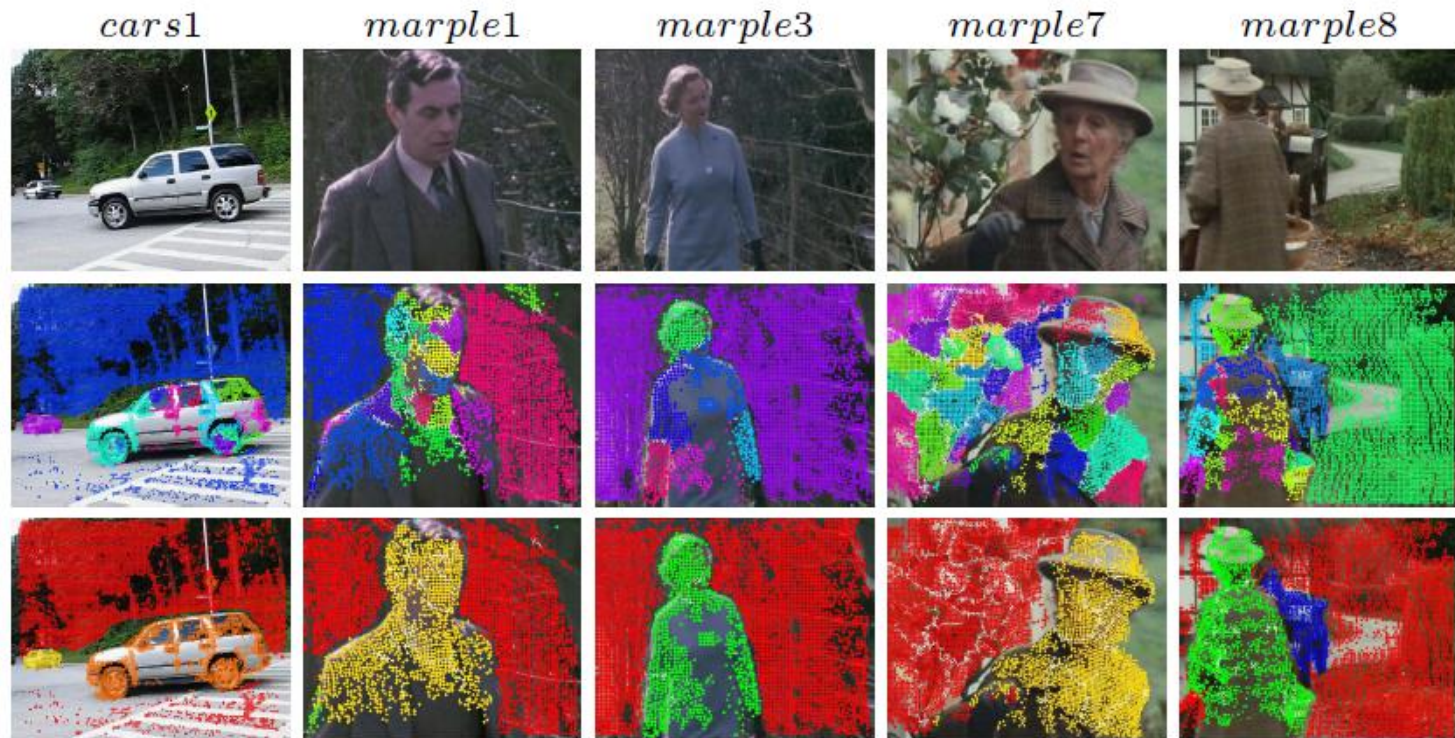



Fig. 6. Motion segmentation results on five sample sequences of the Berkeley Motion Segmentation Dataset [6]. **Second row:** Part segmentation. **Third row:** Object segmentation.

Сегментация 3D сцен, обнаружение 3D объектов


NYU Algorithm on NYU Dataset



NYU1418.jpg

5 most likely categories:


- 0.236223 shoe shop, shoe-shop, shoe store
- 0.027985 confectionery, confectionary
- 0.025233 cinema, movie theater
- 0.024637 butcher shop, meat market
- 0.024317 slot, one-armed bandit



Jia Deng's advisor's advisor's advisor's advisor ←

→ Nevatia & Binford, 1977. → IMAGENET

“These techniques are inadequate for three-dimensional scene analysis for many reasons:”



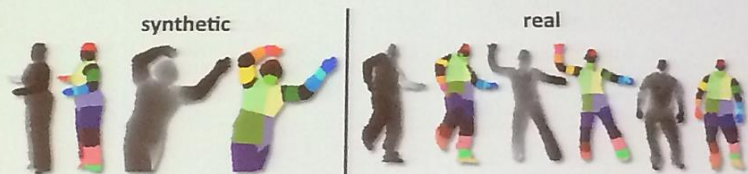
1. Variation
2. Viewpoint
3. Illumination
4. Clutter
5. Occlusion

Sliding Shapes for 3D Object Detection in Depth Images,
Shuran Song, Jianxiong Xiao, ECCV'14

Сегментация 3D сцен, обнаружение 3D объектов

Solution: 3D Depth

- Color Rendering \neq Real Photo
- Depth Rendering \approx Depth from Kinect



Kinect Body Pose Recognition [Shotton et al.]

Sliding Shapes



Input: Kinect Depth Map



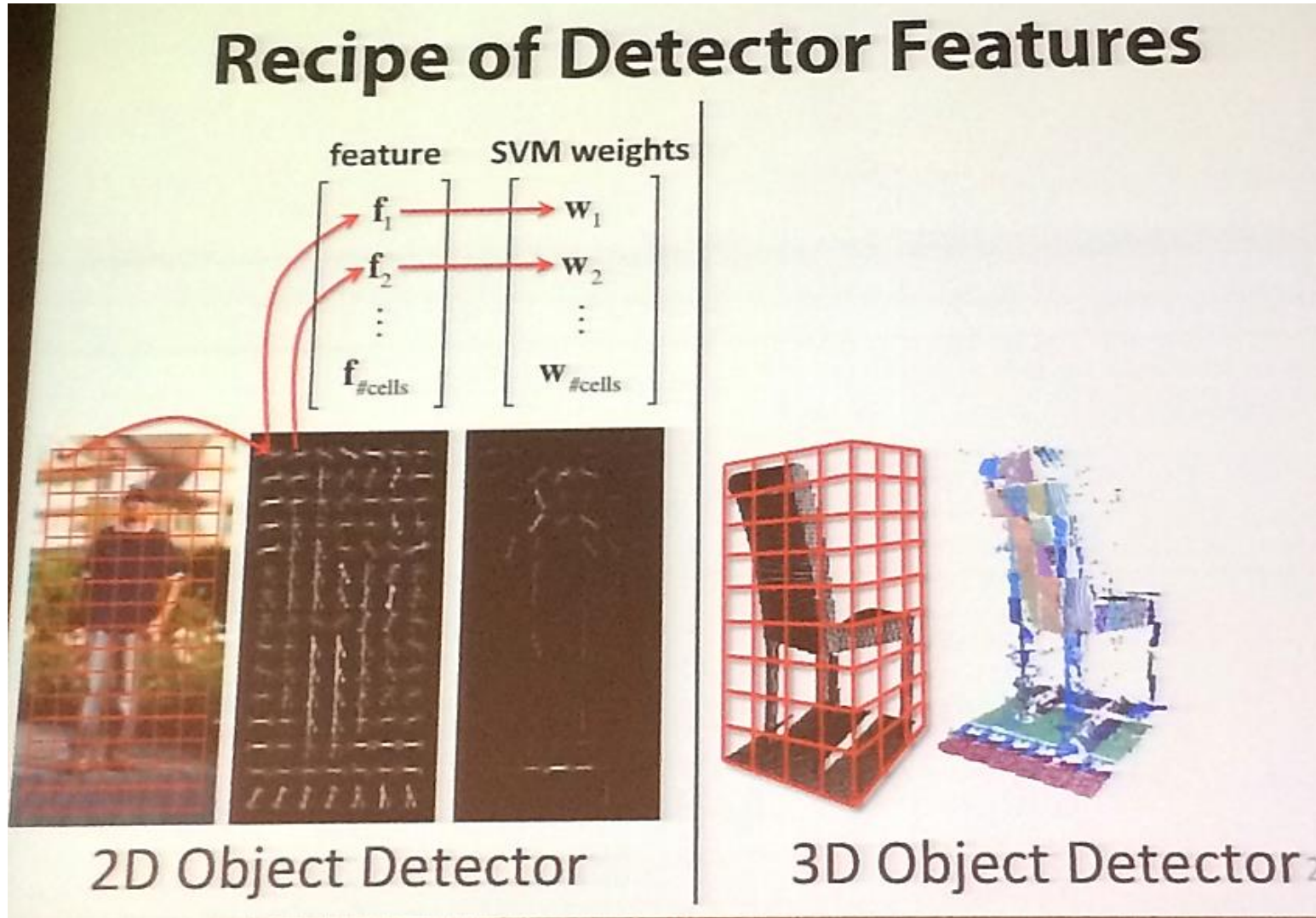
Output: 3D Bounding Box

Code & Data Available

<http://slidingshapes.cs.princeton.edu>

Sliding Shapes for 3D Object Detection in Depth Images,
Shuran Song, Jianxiong Xiao, ECCV'14

Сегментация 3D сцен, обнаружение 3D объектов



Sliding Shapes for 3D Object Detection in Depth Images,
Shuran Song, Jianxiong Xiao, ECCV'14

Сегментация 3D сцен, обнаружение 3D объектов

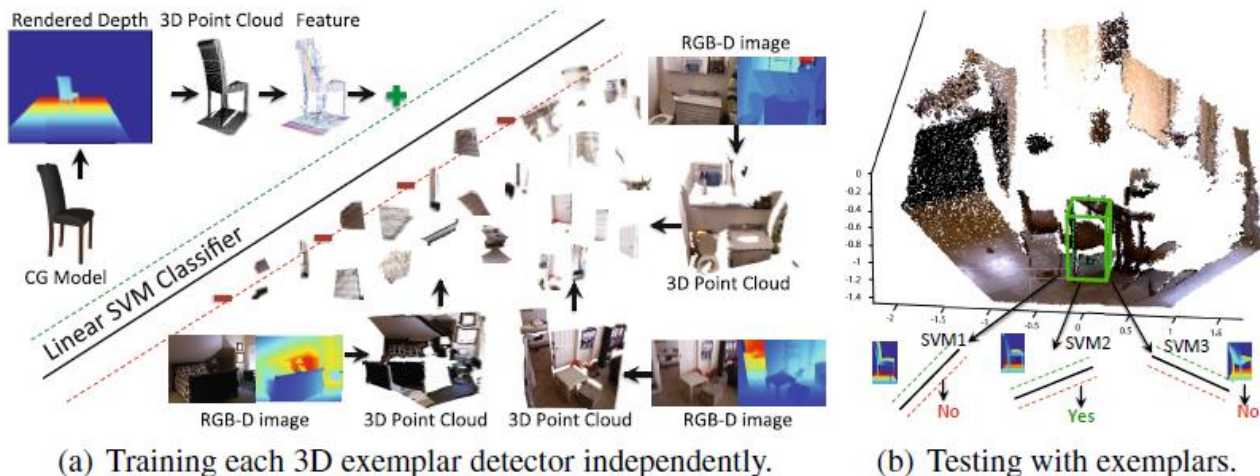
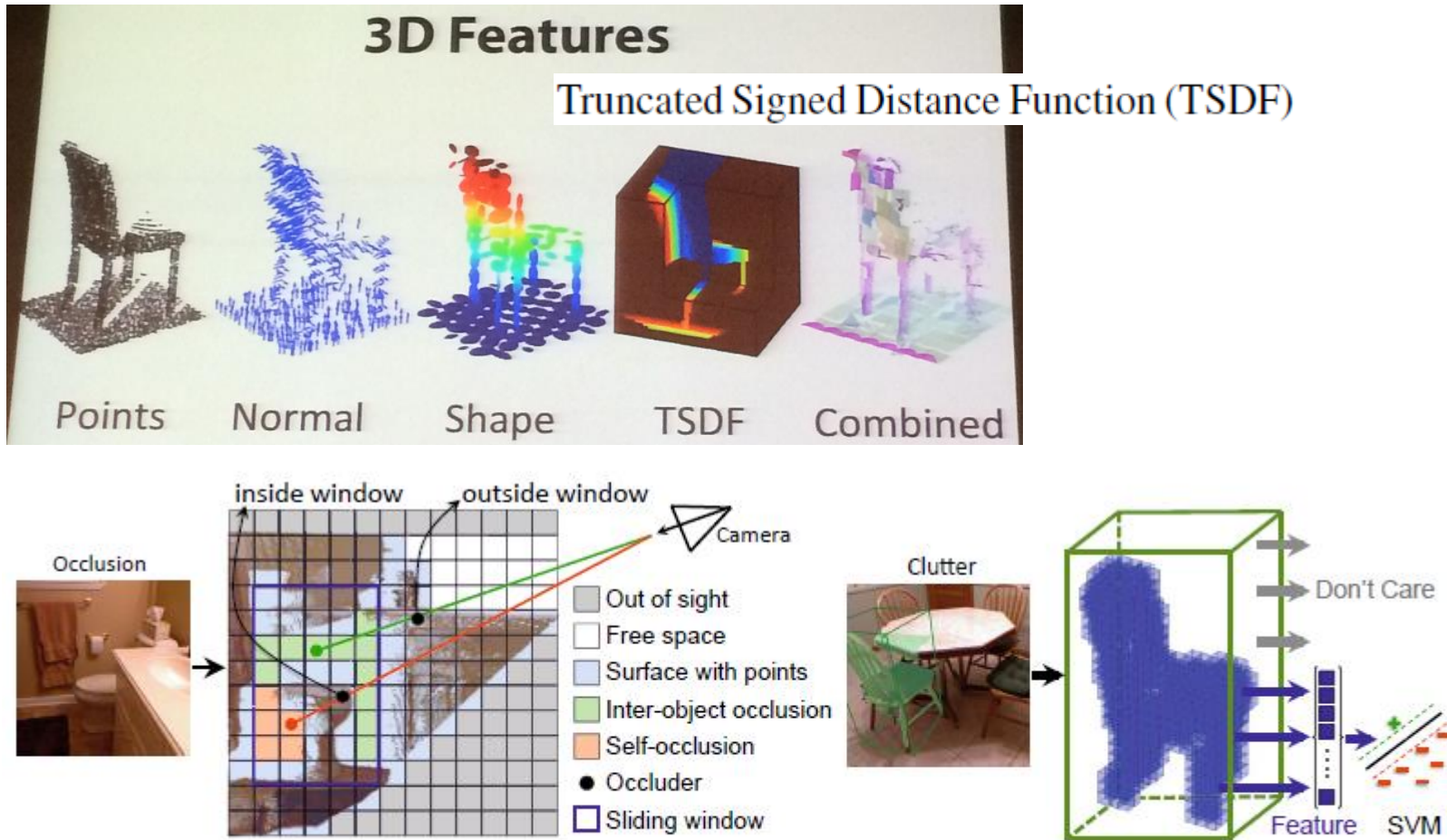


Fig. 1. Sliding Shapes: We extract 3D features of point cloud from depth rendering of CG model to train a 3D classifier. And during testing time, we slide a window in 3D to evaluate the score for each window using an ensemble of Exemplar-SVMs.



Fig. 2. Training procedure: We use a collection of CG models to train a 3D detector. For each CG model, we render it from hundreds of view angles to generate a pool of positive training data. For each rendering, we train an Exemplar-SVM model. And we ensemble all SVMs from renderings of CG chair models to build a 3D chair detector.

Сегментация 3D сцен, обнаружение 3D объектов



(a) Occlusion reasoning using the occluder's location. (b) Occupation mask to slide a shape.

Fig. 4. Beyond sliding windows. Depth and 3D mesh are used to handle occlusion and clutter.

Sliding Shapes for 3D Object Detection in Depth Images,
Shuran Song, Jianxiong Xiao, ECCV'14

Сегментация 3D сцен, обнаружение 3D объектов

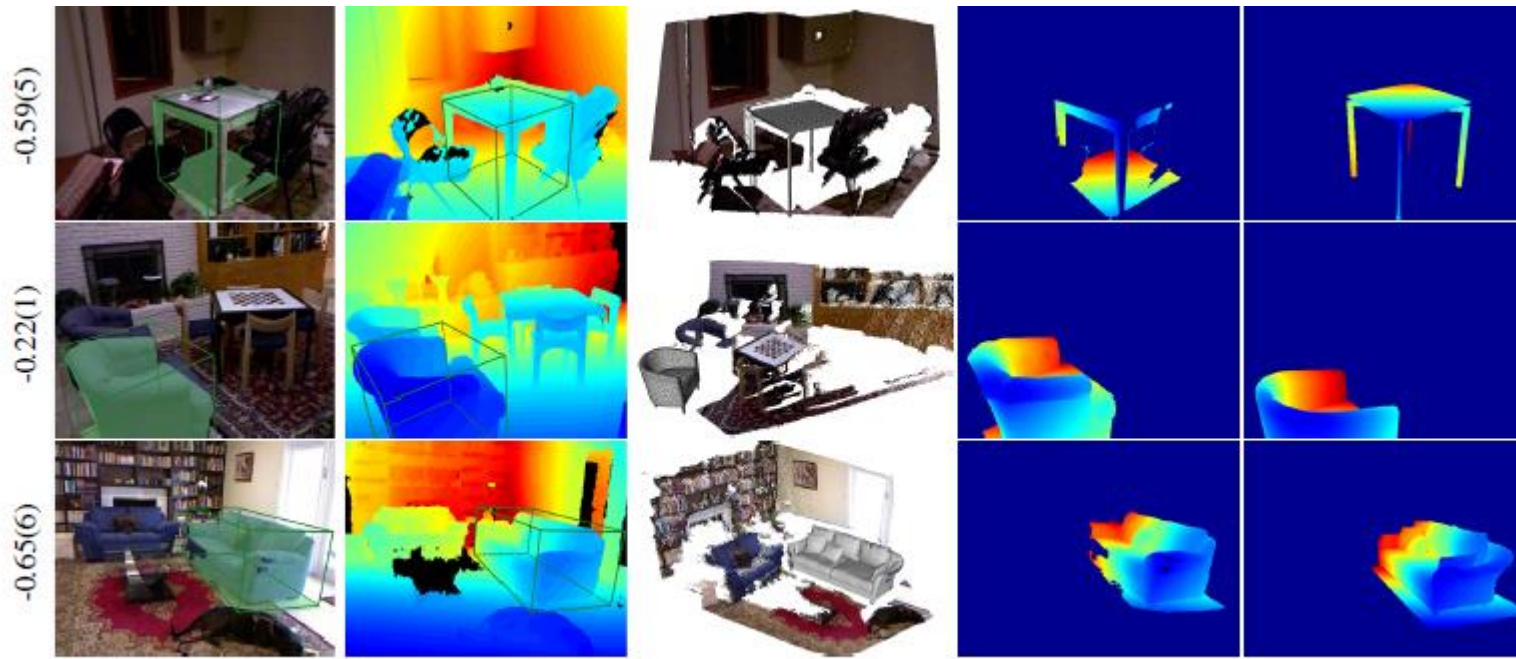
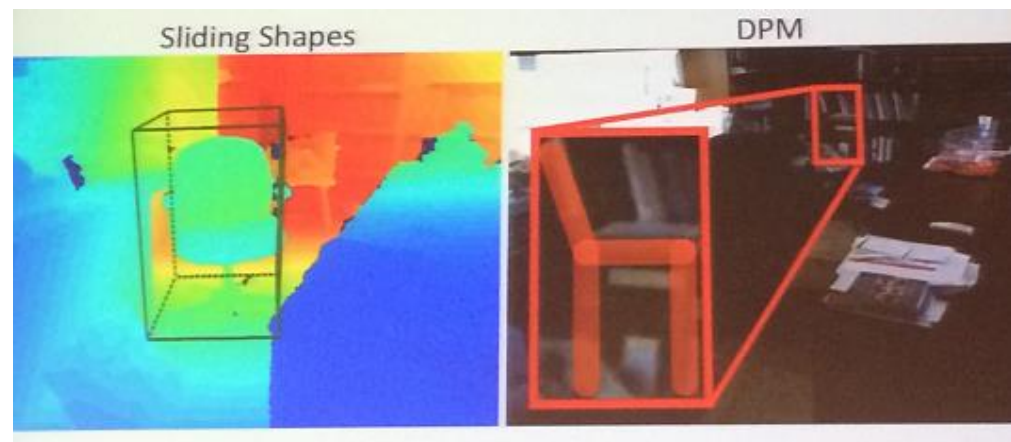
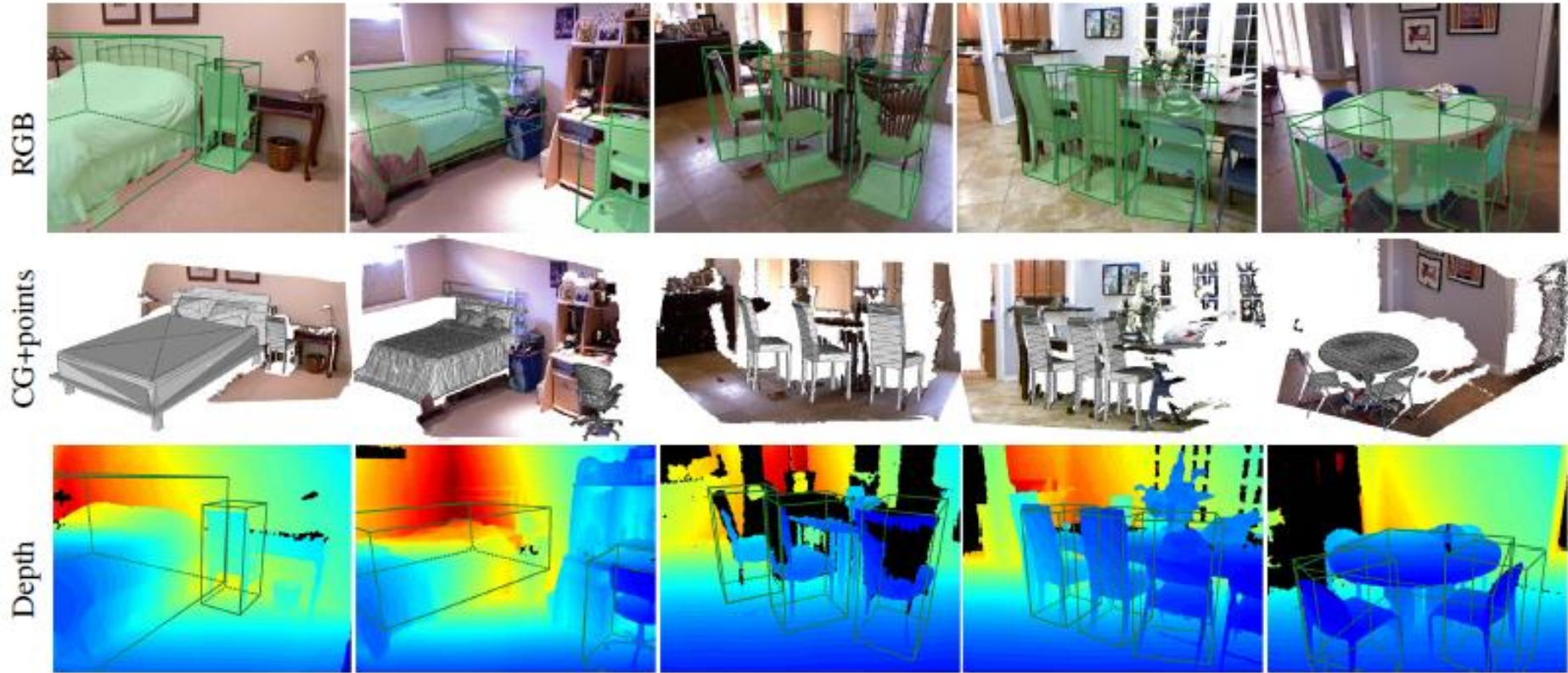


Fig. 6. True positives. Besides labels, our detector also predicts object orientation and 3D model.

Sliding Shapes for 3D Object Detection in Depth Images,
Shuran Song, Jianxiong Xiao, ECCV'14



Сегментация 3D сцен, обнаружение 3D объектов



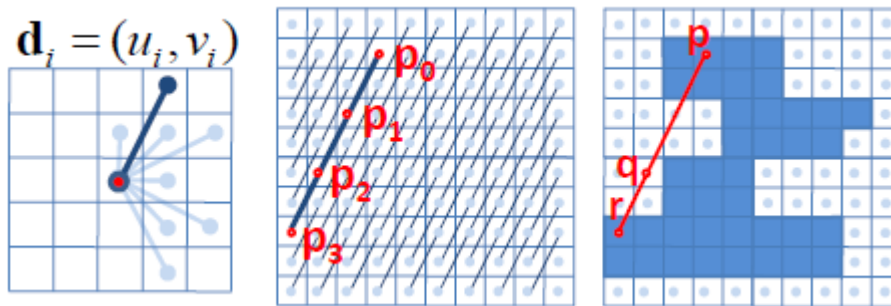
Пример сегментации сцен

Sliding Shapes for 3D Object Detection in Depth Images,
Shuran Song, Jianxiong Xiao, ECCV'14

Сегментация, MRF, Energy-based, Graph-Cut



Fig. 1. Segmentation with convexity shape prior: (a) input image, (b) user scribbles, (c) segmentation with contrast sensitive length regularization. We optimized the weight of length with respect to ground truth. (d) segmentation with convexity shape prior.



$$E_{convexity}(\mathbf{x}) = \sum_{l \in \cup L_i(p,q,r) \in l} \phi(x_p, x_q, x_r).$$

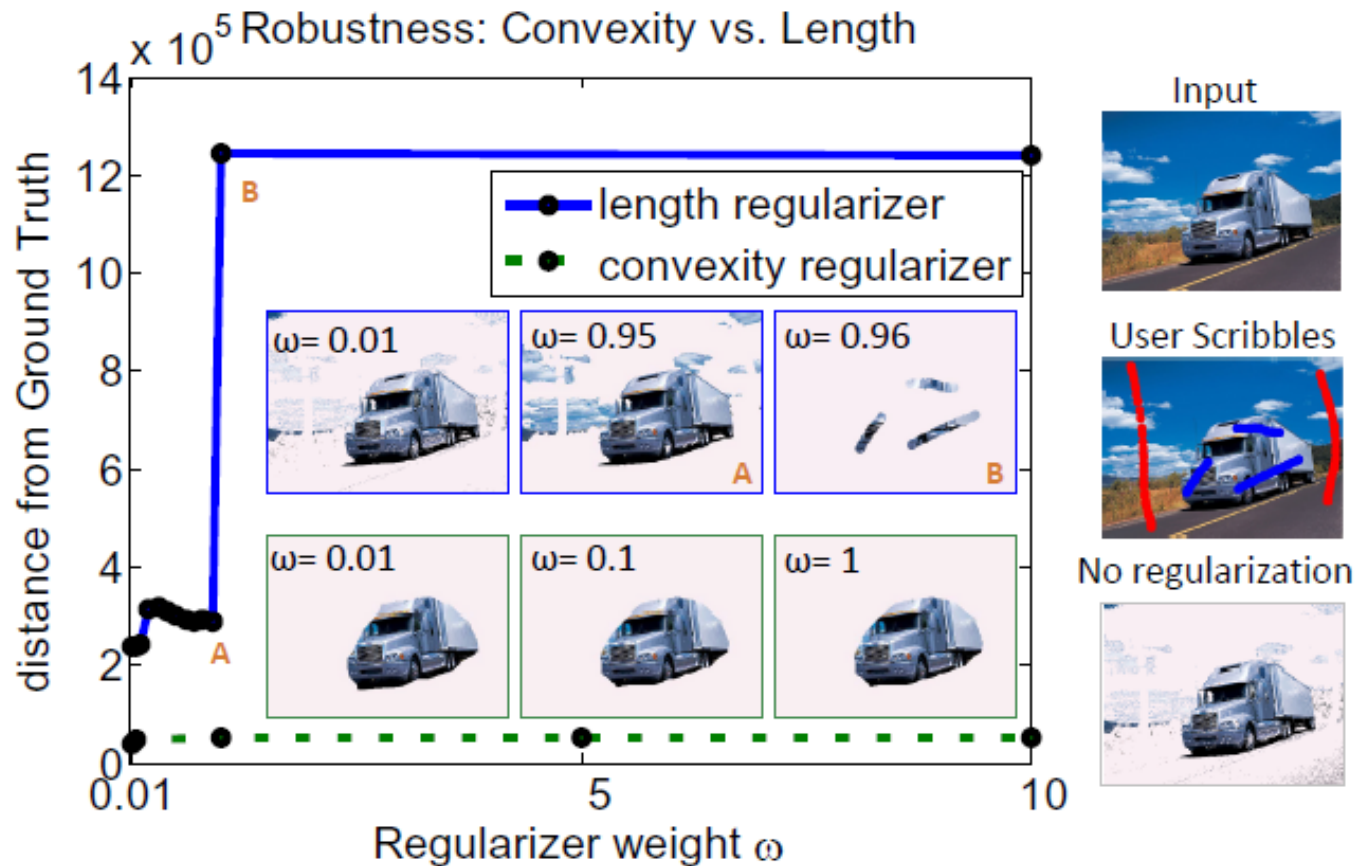
$$E(\mathbf{x}) = E_{convexity}(\mathbf{x}) + E_{sub}(\mathbf{x})$$

$$\phi(x_p, x_q, x_r) = \begin{cases} \infty & \text{if } (x_p, x_q, x_r) = (1, 0, 1) \\ 0 & \text{otherwise.} \end{cases}$$

Convexity Shape Prior for Segmentation,

Lena Gorelick, Olga Veksler, **Yuri Boykov**, and Claudia Nieuwenhuis, ECCV'14

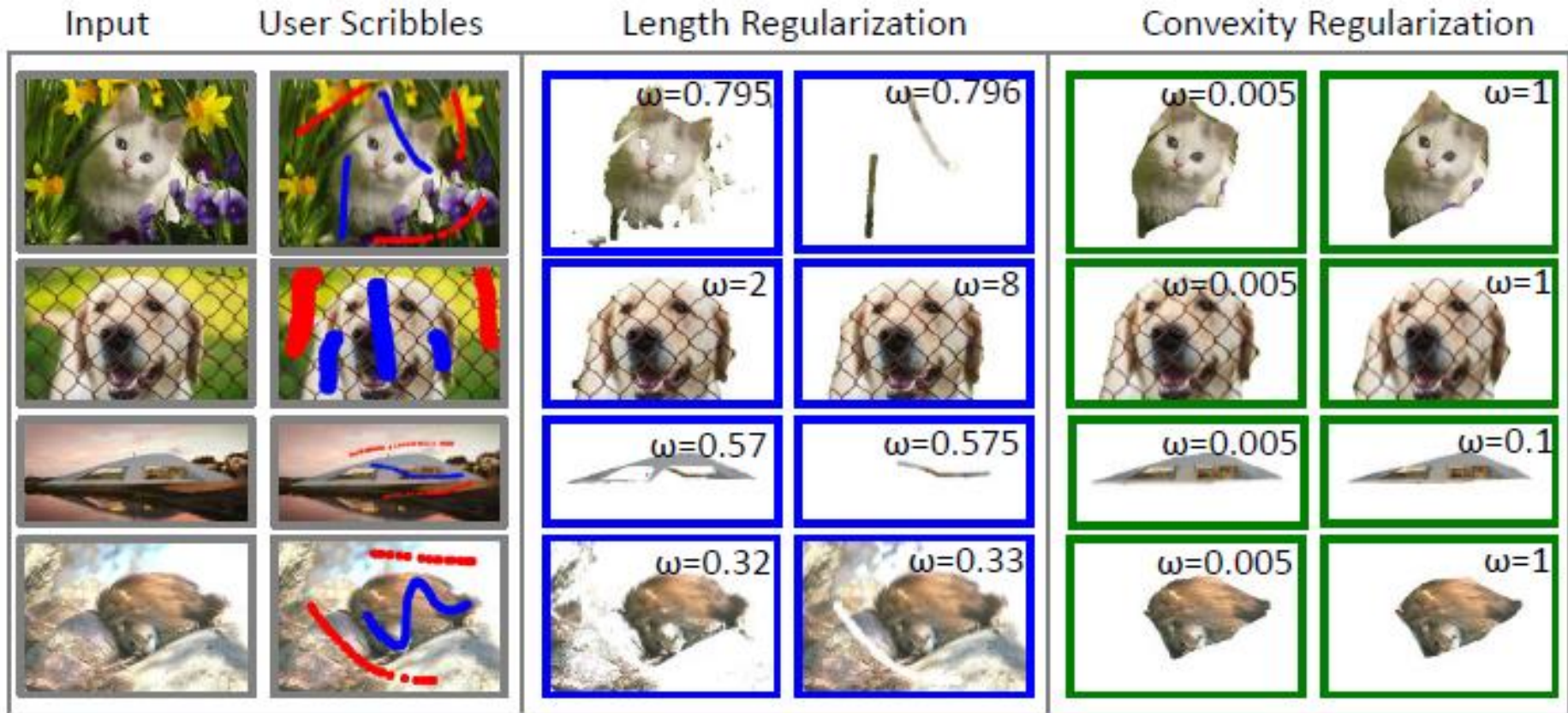
Сегментация, MRF, Energy-based, Graph-Cut



Convexity Shape Prior for Segmentation,

Lena Gorelick, Olga Veksler, **Yuri Boykov**, and Claudia Nieuwenhuis, ECCV'14

Сегментация, MRF, Energy-based, Graph-Cut

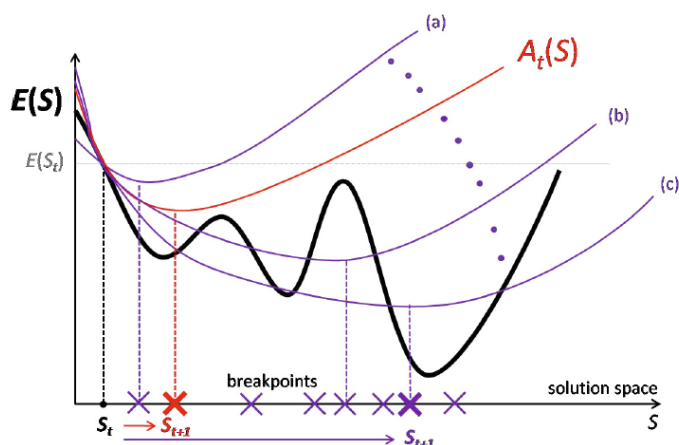


Convexity Shape Prior for Segmentation,

Lena Gorelick, Olga Veksler, **Yuri Boykov**, and Claudia Nieuwenhuis, ECCV'14

Сегментация, MRF, Energy-based, Graph-Cut

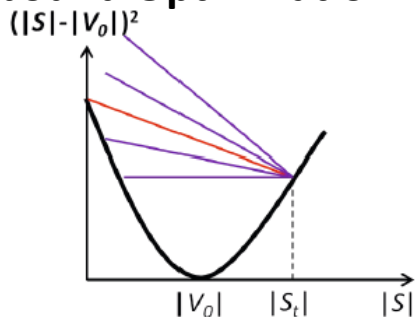
Bound Optimization



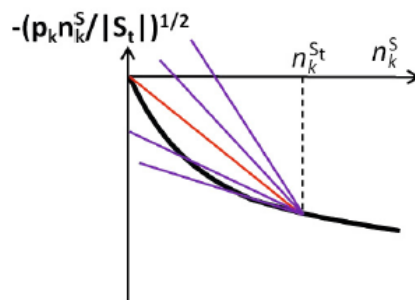
Algorithm 1. PARAMETRIC PSEUDO-BOUND CUTS (PPBC)

- 1 $S_0 \leftarrow S_{init}$
 - 2 For $t = 0, 1, 2, \dots$, repeat until convergence
 - 3 Construct an auxiliary function $A_t(S)$ at current solution S_t ;
 - 4 Combine $A_t(S)$ with unary relaxation term $R_t(S)$ to form pseudo-bound $\mathcal{F}_t(S, \lambda) = A_t(S) + \lambda R_t(S)$
 - 5 //Optimize the parametric family of pseudo-bounds $S^\lambda = \arg \min_S \mathcal{F}_t(S, \lambda)$, for $\lambda \in \Lambda$
 - 6 //Score candidate solutions and update $\lambda^* = \arg \min_\lambda E(S^\lambda)$, $S_{t+1} \leftarrow S^{\lambda^*}$
-

Pseudo-bound Optimization



(a) Volumetric prior, Sec.3.1.1



(b) Bhattacharyya prior, Sec.3.1.3

Fig. 3. Pseudo-bound families for two cardinality functions. Auxiliary functions are red.

Pseudo-bound Optimization for Binary Energies,

Meng Tang, Ismail Ben Ayed, and **Yuri Boykov**, ECCV'14

Сегментация, MRF, Energy-based, Graph-Cut

Non-submodular Pairwise Energies

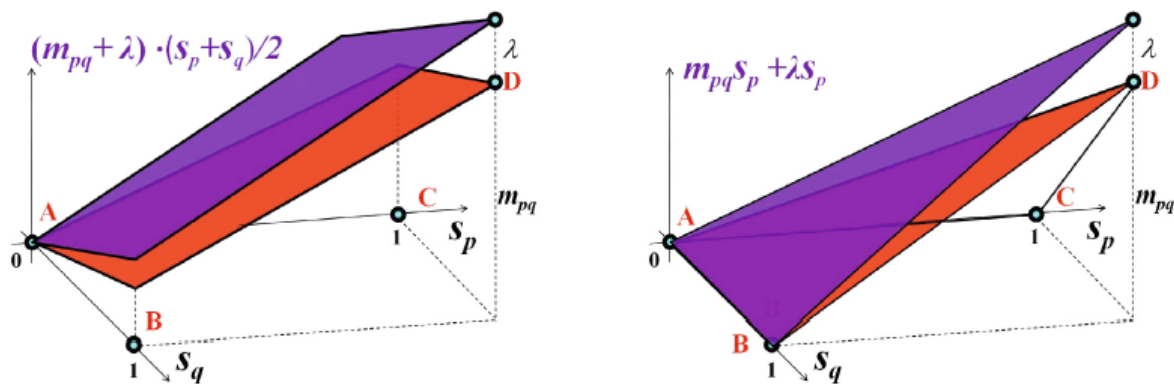


Fig. 4. pPBC-T: Pseudo-bounds (purple) and auxiliary functions (red) of non-submodular potential $m_{pq}s_p s_q$ for current configuration $s_{p,t} = 0, s_{q,t} = 0$ (left) and $s_{p,t} = 0, s_{q,t} = 1$ (right).

Table 1. Auxiliary functions [11] and weighted bound relaxation term for pPBC-T

$(s_{p,t}, s_{q,t})$	Auxiliary function	relaxation term (pPBC-T)
(0, 0)	$m_{pq}(s_p + s_q)/2$	$\lambda(s_p - s_{p,t} + s_q - s_{q,t})$
(0, 1)	$m_{pq}s_p$	$\lambda(s_p - s_{p,t})$
(1, 0)	$m_{pq}s_q$	$\lambda(s_q - s_{q,t})$
(1, 1)	$m_{pq}(s_p + s_q)/2$	$\lambda(s_p - s_{p,t} + s_q - s_{q,t})$

Pseudo-bound Optimization for Binary Energies,

Meng Tang, Ismail Ben Ayed, and **Yuri Boykov**, ECCV'14

Сегментация, MRF, Energy-based, Graph-Cut

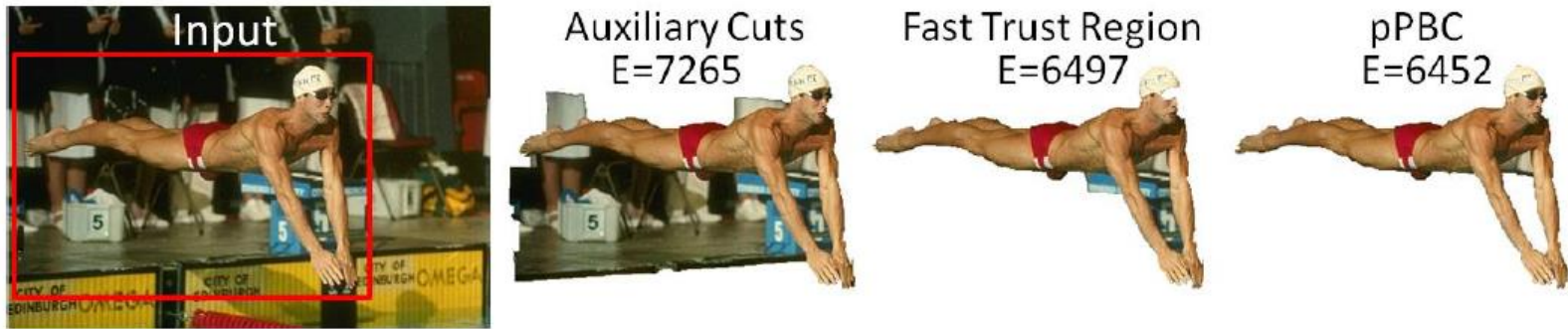


Fig. 2. Matching target foreground color distribution using auxiliary cuts [3], fast trust region [12] and pPBC. pPBC achieves the lowest energy.

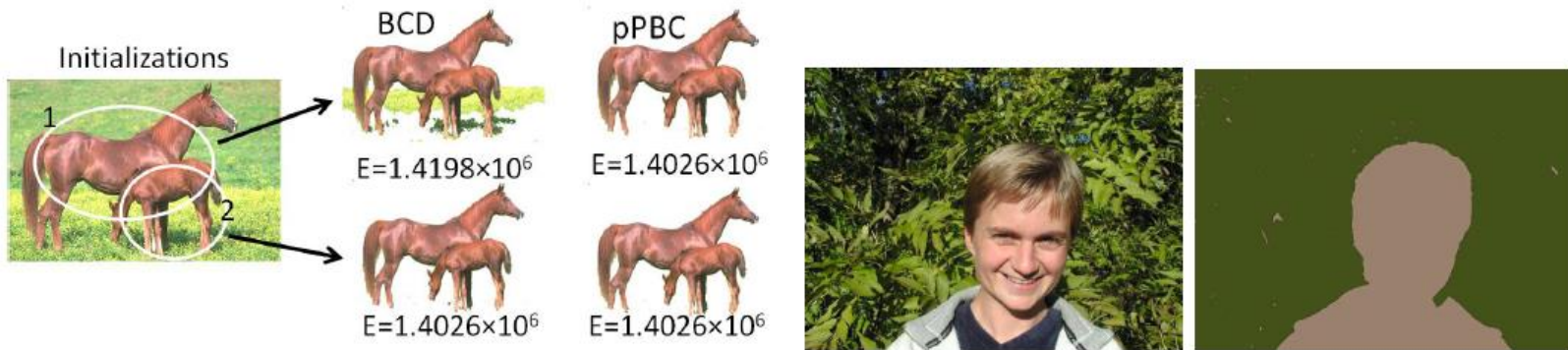
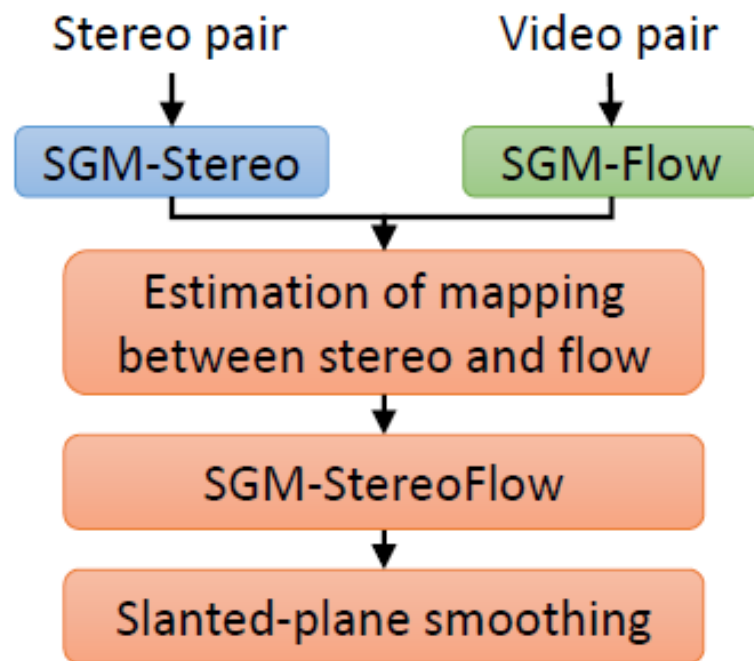


Fig. 5. Left: interactive segmentations with BCD (GrabCut) or pPBC from different initialization (ellipses). Proposed pPBC method is more robust to inferior initialization. Right: unsupervised figure-ground segmentation with pPBC. Average color is shown.

MRF, Segmentation, 3D Flow

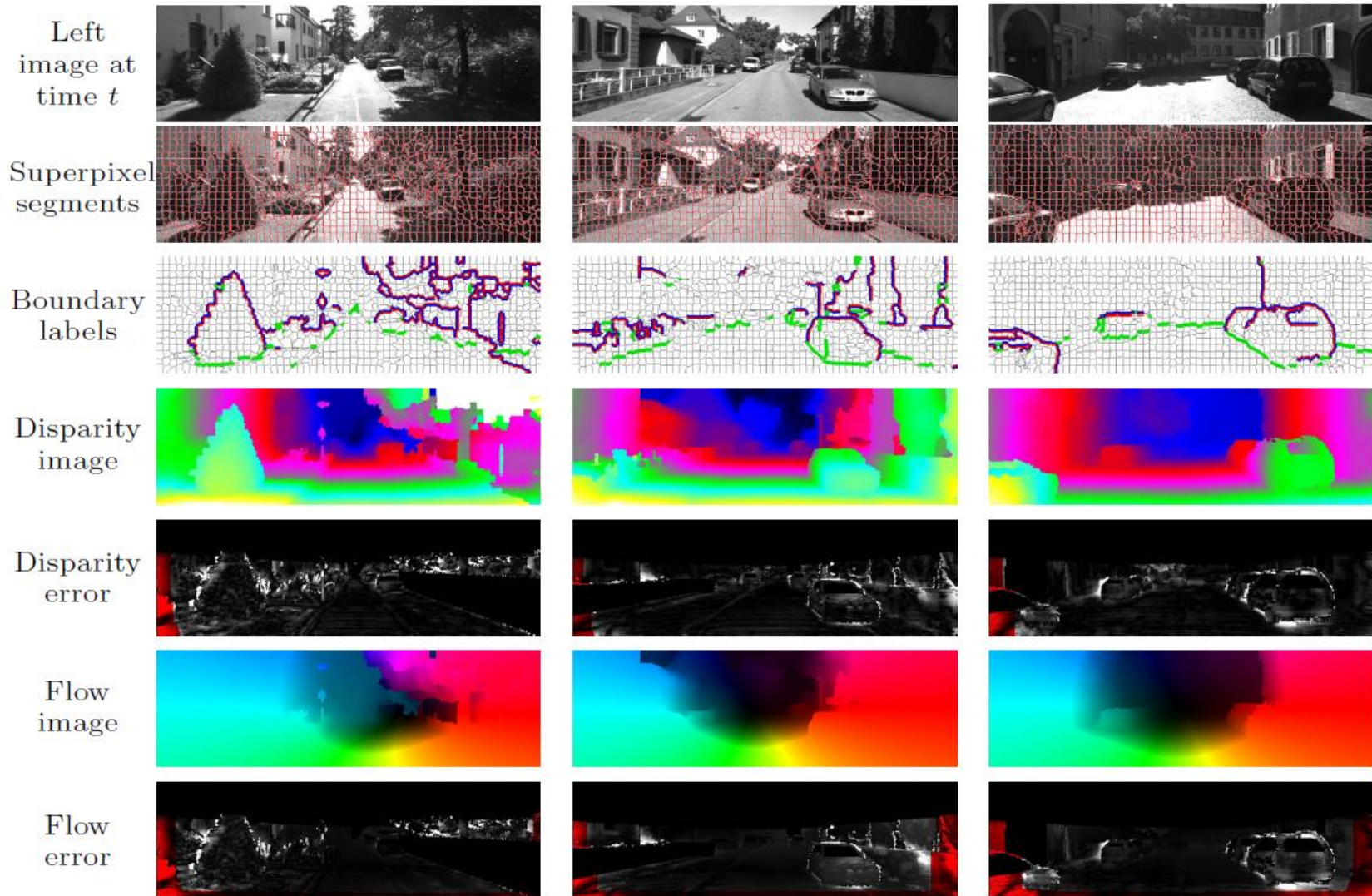
$$\begin{aligned} E(s, \theta, f, o, \mathcal{I}, d) = & \underbrace{\sum_{\mathbf{p}} E_{\text{col}}(\mathbf{p}, c_{s_p})}_{\text{color-data}} + \underbrace{\lambda_{\text{pos}} \sum_{\mathbf{p}} E_{\text{pos}}(\mathbf{p}, \mu_{s_p})}_{\text{location}} + \underbrace{\lambda_{\text{depth}} \sum_{\mathbf{p}} E_{\text{depth}}(\mathbf{p}, \theta_{s_p}, f_p)}_{\text{depth-data}} \\ & + \underbrace{\lambda_{\text{smo}} \sum_{\{i,j\} \in \mathcal{N}_{\text{seg}}} E_{\text{smo}}(\theta_i, \theta_j, o_{i,j})}_{\text{plane-smoothness}} + \underbrace{\lambda_{\text{com}} \sum_{\{i,j\} \in \mathcal{N}_{\text{seg}}} E_{\text{prior}}(o_{i,j})}_{\text{label-prior}} \\ & + \underbrace{\lambda_{\text{bou}} \sum_{\{\mathbf{p}, \mathbf{q}\} \in \mathcal{N}_8} E_{\text{bou}}(s_p, s_q)}_{\text{boundary-length}} \end{aligned}$$



Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation

Koichiro Yamaguchi, David McAllester, and **Raquel Urtasun**, ECCV'14

MRF, Segmentation, 3D Flow



Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation

Koichiro Yamaguchi, David McAllester, and Raquel Urtasun, ECCV'14

Face, 3D Reconstruction, 3D Flow

Input

Video



OR

Still images



Internet photos (same person)



Method Overview

Average Shape



Pose Estimation



3D Flow



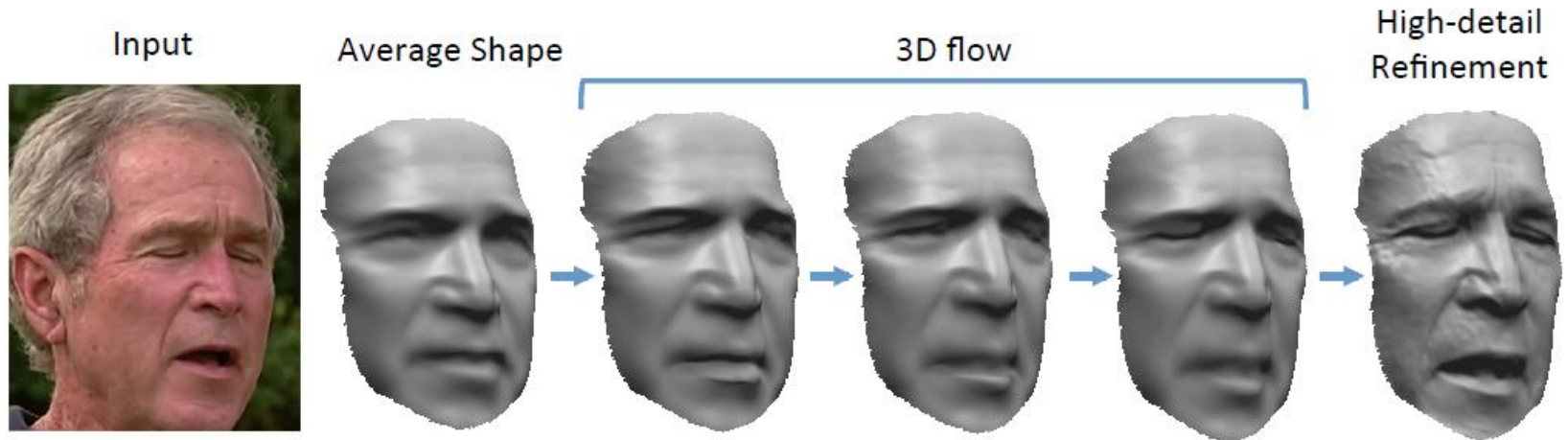
High Detail Refinement



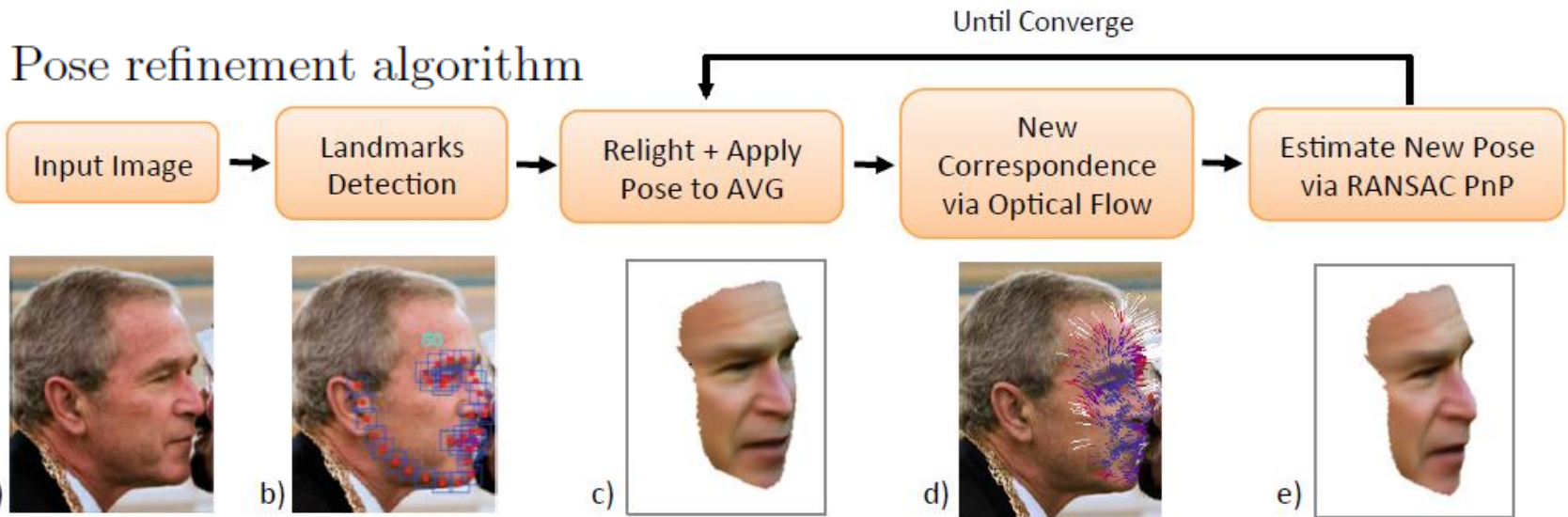
Total Moving Face Reconstruction

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz, ECCV'14

Face, 3D Reconstruction, 3D Flow



3D flow convergence example



Total Moving Face Reconstruction

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz, ECCV'14

Face, 3D Reconstruction, 3D Flow



Total Moving Face Reconstruction

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz, ECCV'14

Face, 3D Reconstruction, 3D Flow



Total Moving Face Reconstruction

Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz, ECCV'14

3D Reconstruction, 4D Segmentation, 3D Flow

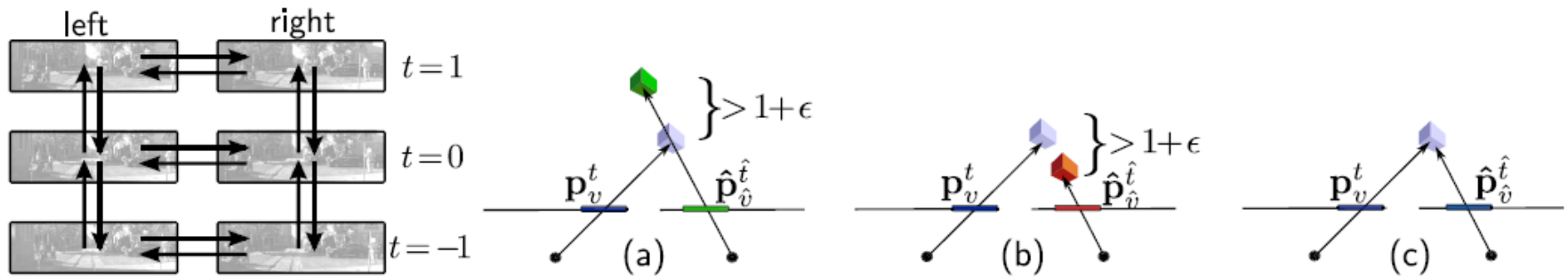


Fig. 2. (left) Data terms in the three-frame case: Consistency is enforced for spatial and direct temporal neighbors (black arrows). (right) Illustration of the per pixel data term: (a) impossible case, (b) occlusion (c) normal case (see text for more details.)

$$E(\mathcal{P}, \mathcal{S}) = E_D(\mathcal{P}, \mathcal{S}) + \lambda E_R(\mathcal{P}, \mathcal{S}) + \mu E_S(\mathcal{S}).$$

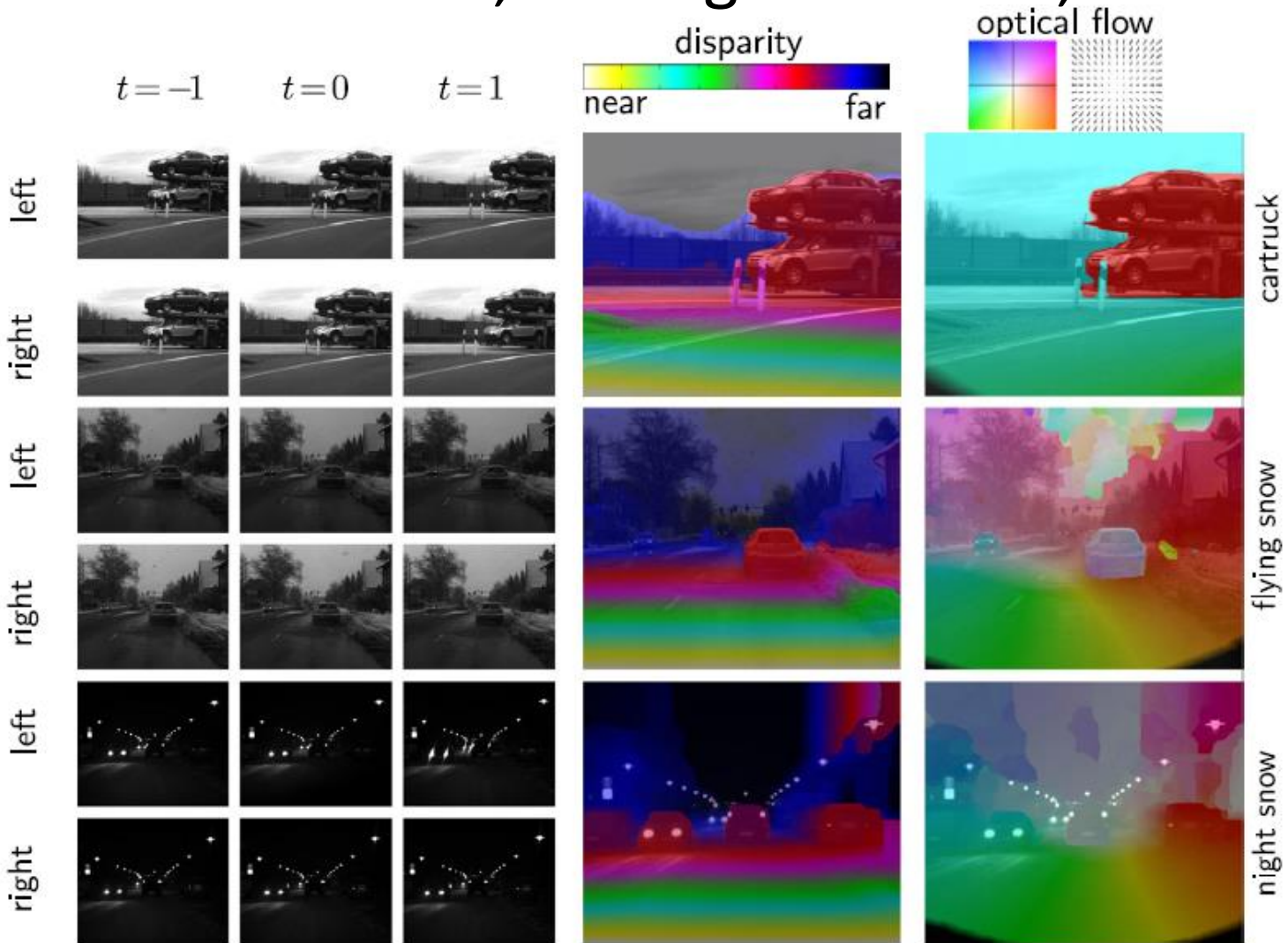
View-Consistent Data Term

Shape and Motion Regularization

Spatial Segmentation Regularization

View-Consistent 3D Scene Flow Estimation over Multiple Frames,
Christoph Vogel, Stefan Roth, and **Konrad Schindler**, ECCV'14

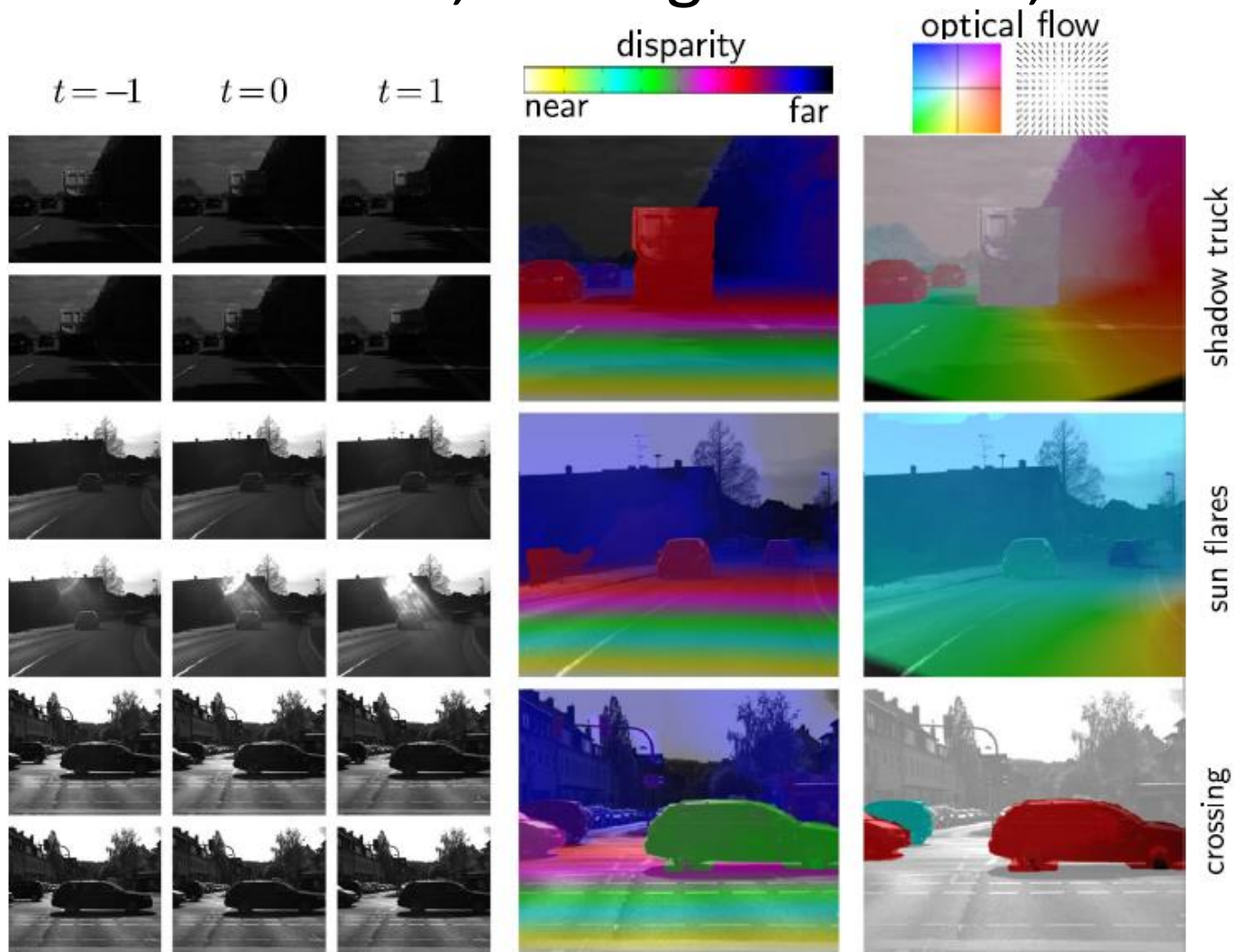
3D Reconstruction, 4D Segmentation, 3D Flow



(left) Input frames. (right) Reconstructed scene flow, reprojected to disparity and 2D flow field.

View-Consistent 3D Scene Flow Estimation over Multiple Frames,
Christoph Vogel, Stefan Roth, and Konrad Schindler, ECCV'14

3D Reconstruction, 4D Segmentation, 3D Flow



(left) Input frames. (right) Reconstructed scene flow, reprojected to disparity and 2D flow field.

View-Consistent 3D Scene Flow Estimation over Multiple Frames,
Christoph Vogel, Stefan Roth, and Konrad Schindler, ECCV'14

Анализ формы:

Morphology,
Shape Analysis,
Manifolds, SPD

Морфологическая фильтрация изображений

Tubular Structure Filtering by Ranking Orientation Responses of Path Operators



UNIVERSITÉ
— PARIS-EST

Odyssée Merveille^{1,2}, Hugues Talbot¹, Laurent Najman¹ and Nicolas Passat²
¹ Université Paris-Est, LIGM, UPEMLV-ESIEE-CNRS, France
² Université de Reims Champagne-Ardenne, CRSTIC, France



Introduction

Tubular objects, like vascular networks or fibres in materials science, have been of interest for some time in computer vision. Usually, tubular structure filtering uses an analysis of the three principal directions of the Hessian which is a local feature. We propose a low-level tubular structure detection filter based on paths, which are semi global features that avoid any blurring effect induced by scale-space convolution.

Context

3D Path Operators can filter thin objects, which means both tubular and plane-like structures. Our strategy for filtering only tubular structures derives from the simple observation of figure 1. A blob, a plane and a tubular structure in 3D can be distinguished by "counting" the number of responses of any oriented filter as path operators.

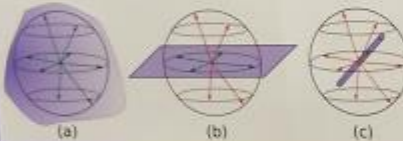
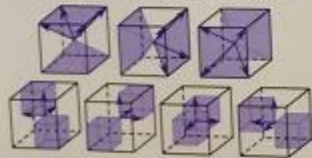


Figure 1: When sampling orientations from a point, in an isotropic structure a.k.a a blob (a), oriented operators all respond nearly identically (green arrow). In a plane (b), some proportion respond positively. In a tube (c), only a few orientations respond.

Ranking Orientation Responses of Path Operators: RORPO

Hypothesis: Plane structures are detected in at least one more RPO orientation than tubular structures.

Figure 2: The seven 3D orientations

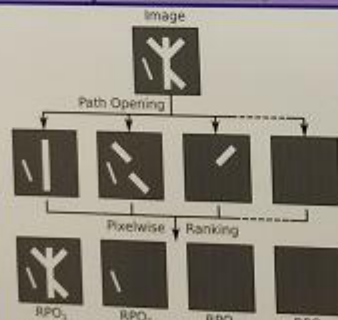


Based on the orientations of figure 2, we proposed the RORPO operator:

$$RORPO = RPO_1 - RPO_2$$

with RPO_1 : 1st ranked orientation (=RPO result)
 RPO_2 : 2nd ranked orientation

- We showed on synthetic images that RPO, with these orientations detects:
- Tubular structures in at most 3 orientations
 - Plane structures in at least 5 orientations



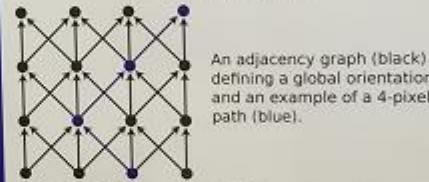
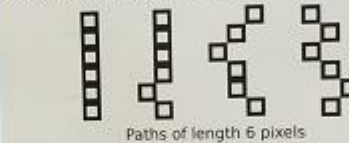
A structure present in 2 RPO orientations will appear in RPO1 and RPO2.
 So according to this hypothesis we chose $i=4$:

$$RORPO = RPO_1 - RPO_4$$

- RPO_1 contains all tubes and planes
- RPO_4 contains no tubes and all planes
- RORPO contains all tubes and no planes

Previous Work: Path Operators

Definition: Morphological opening or closing which use as structuring element a set of oriented connected pixels of fixed length called paths.



Robust Path Opening (RPO): A version of Path Opening that is robust to noise.



Filtering tubular structures in all directions requires the fusion of RPO in several orientations.



Result of a 2D RPO (b) on an initial image (a)

Results and Comparisons

Synthetic image



MIP of the initial image (a) and the RORPO result (b)

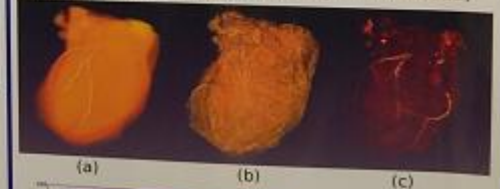
Brain MRI



Volume Rendering of the initial image (a) and the RORPO result (b)

Comparison with Frangi's Vesselness on Heart CT

Quantitative comparison with Frangi's Vesselness (gold standard in tubular filtering) on 15 patients of the Rotterdam repository (Challenge MICCAI 2012)



(a) Initial image (b) Frangi's result (c) RORPO result



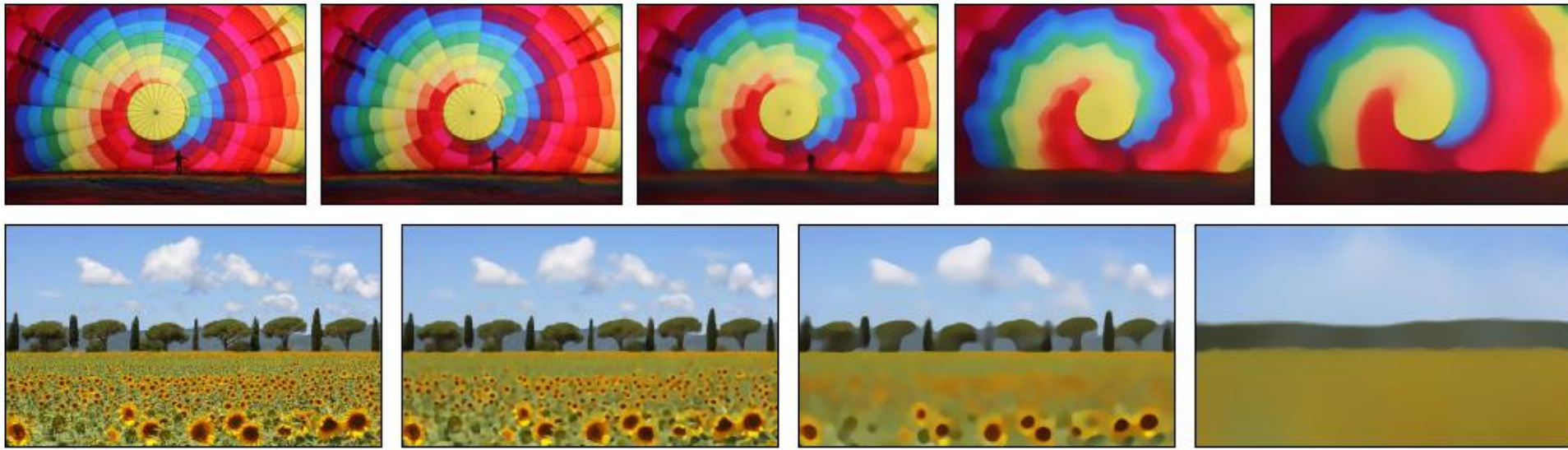
Volume Rendering of (a) Initial image and ground truth (b) Frangi's result (c) RORPO result

(d) ROC curves on 15 Patients

Морфологическая фильтрация изображений

Our main algorithm has only **1 line** of code

```
while(iter-->0) res = bilateralFilter(im,res,sc,sm);
```



Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Морфологическая фильтрация изображений

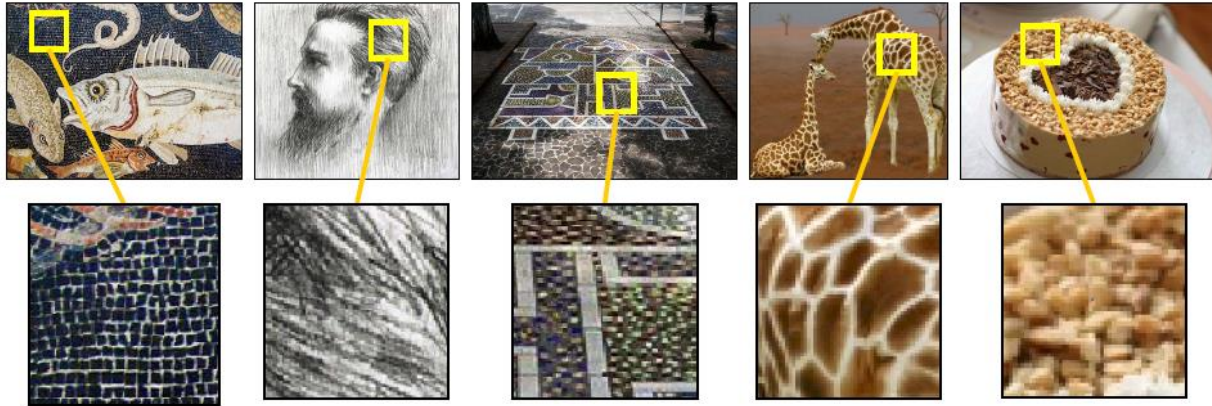


Fig. 1. Examples of high-contrast details in natural images. As explained, edge-aware filters aim to maintain them due to the large magnitude of edges.

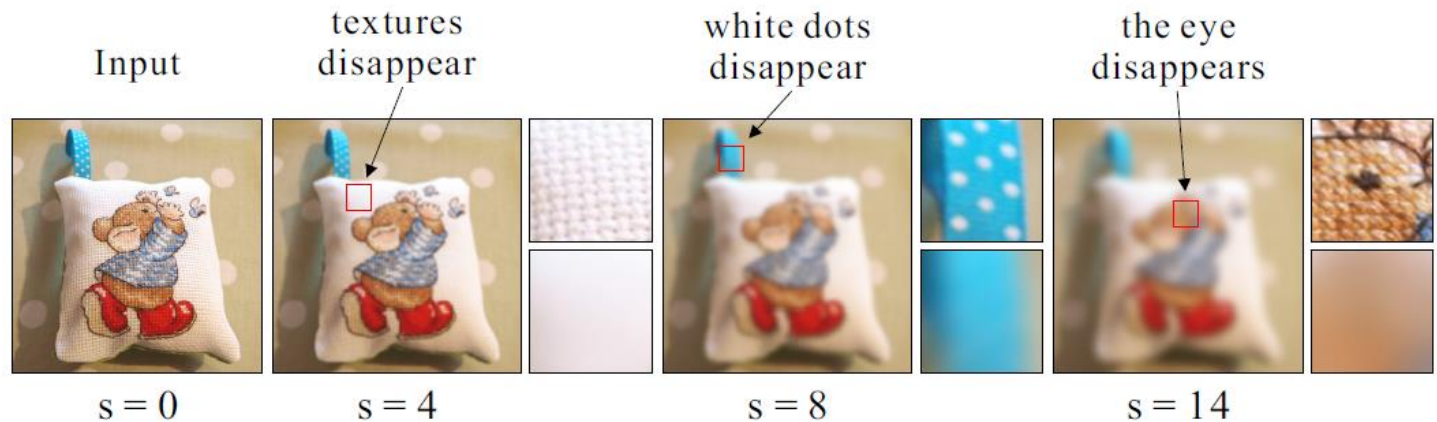


Fig. 2. Illustration of scales. As the Gaussian kernel gets larger, more and more structures disappear.

Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Морфологическая фильтрация изображений

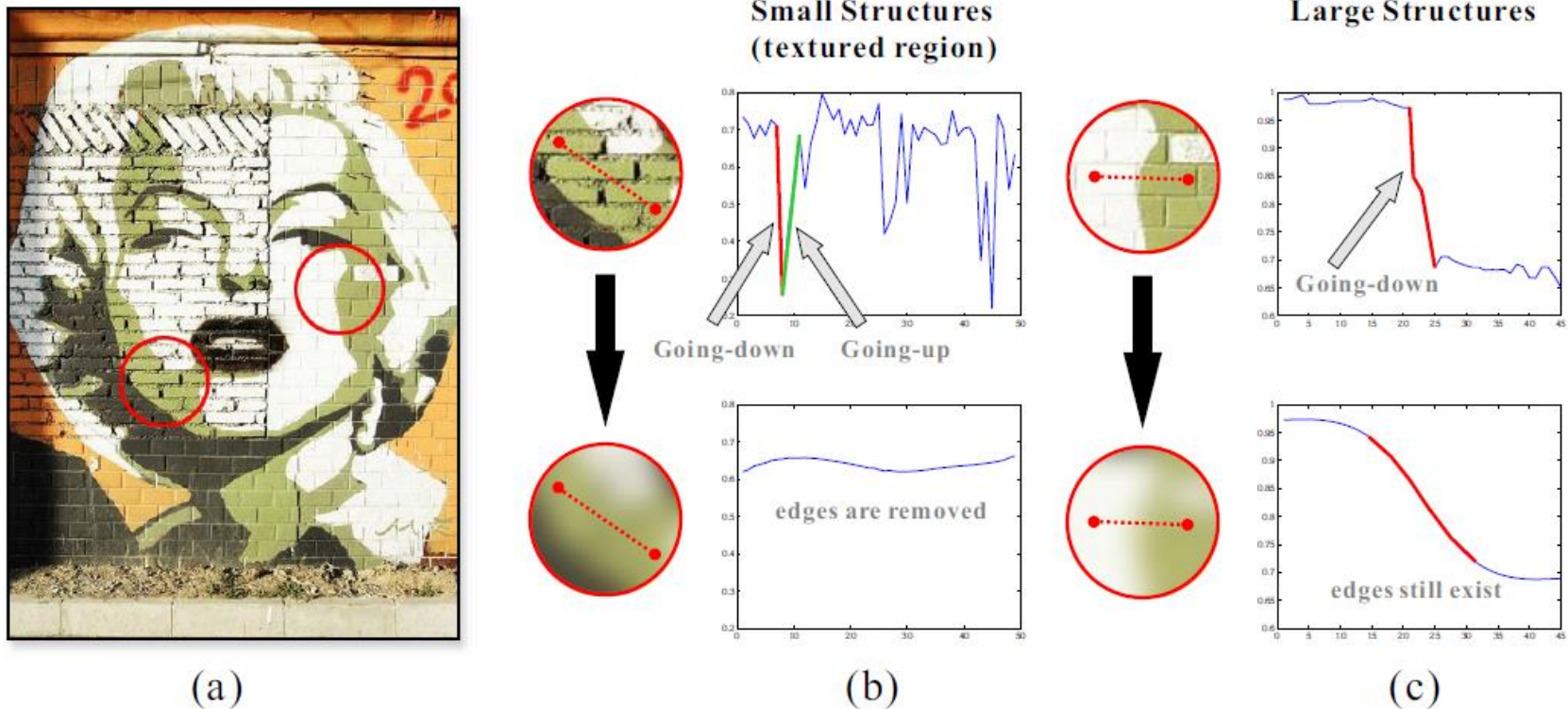


Fig. 3. Comparison of small- and large-structure results after Gaussian filtering. (a) Input image. (b)-(c) 1D signals of pixel values in two lines. The upper signals are the input and the lower ones are results of Gaussian filtering.

Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Морфологическая фильтрация изображений

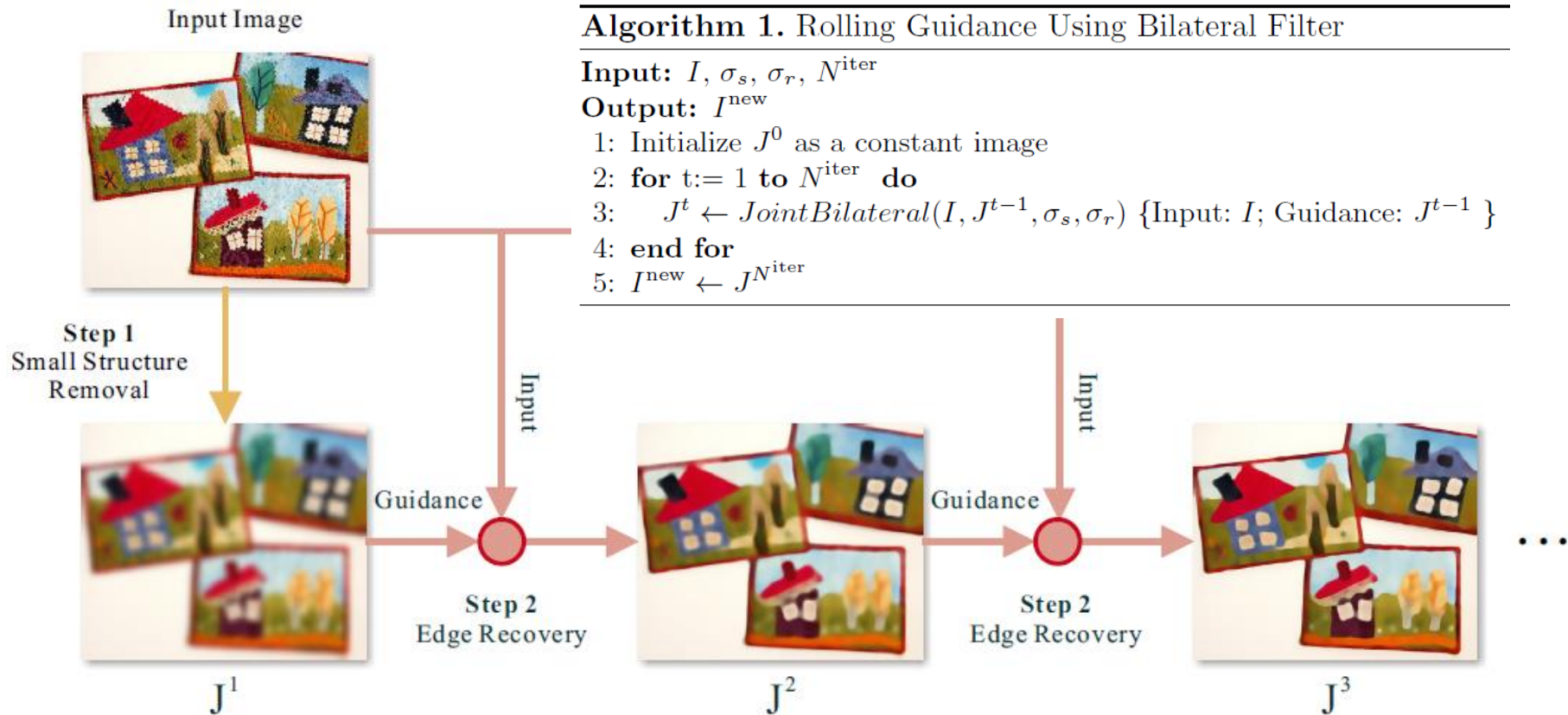


Fig. 4. Flow chart of our method. It contains two steps respectively for small structure removal (Section 4.1) and edge recovery (Section 4.2). Step 2 is an iterative process. The final result is obtained in 3–5 iterations.

Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Морфологическая фильтрация изображений

Edge Recovery

The iterative edge recovery step forms the major contribution in our method. In this process, an image J is iteratively updated. We denote J^{t+1} as the result in the t -th iteration. Initially, J^1 is set as G in Eq. (2), which is the output of Gaussian filtering. The value of J^{t+1} in the t -th iteration is obtained in a joint bilateral filtering form given the input I and the value in previous iteration J^t :

$$J^{t+1}(p) = \frac{1}{K_p} \sum_{q \in N(p)} \exp \left(-\frac{\|p - q\|^2}{2\sigma_s^2} - \frac{\|J^t(p) - J^t(q)\|^2}{2\sigma_r^2} \right) I(q),$$

where

$$K_p = \sum_{q \in N(p)} \exp \left(-\frac{\|p - q\|^2}{2\sigma_s^2} - \frac{\|J^t(p) - J^t(q)\|^2}{2\sigma_r^2} \right)$$

for normalization. I is the same input image used in Eq. (2). σ_s and σ_r control the spatial and range weights respectively.

He, K., Sun, J., Tang, X.: Guided image filtering. ECCV 2010

Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Морфологическая фильтрация изображений

Per-pixel Intensity Difference Between Two Iterations

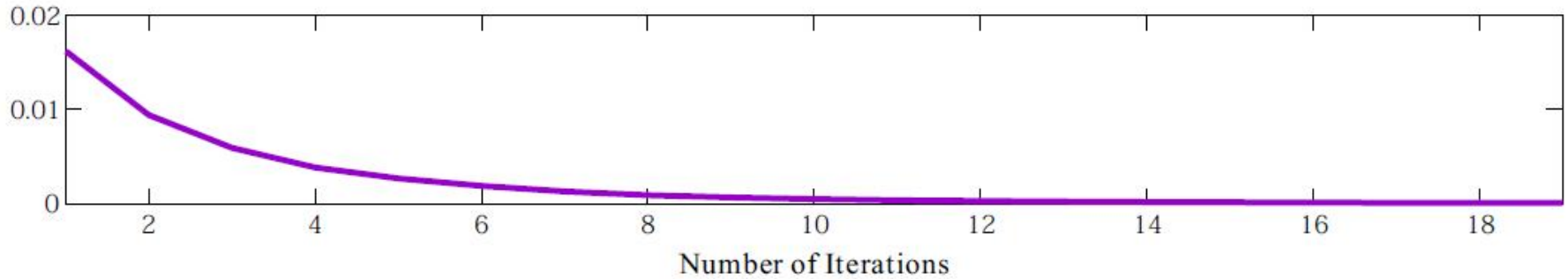


Fig. 7. Plot of difference between input and output images in iterations. The difference of two successive iterations reduces monotonically and the result is guaranteed not an all-constant image. We use $\sigma_s = 4$ and $\sigma_r = 0.1$ for this example. Please view them in the original resolutions to compare all details.

Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Морфологическая фильтрация изображений

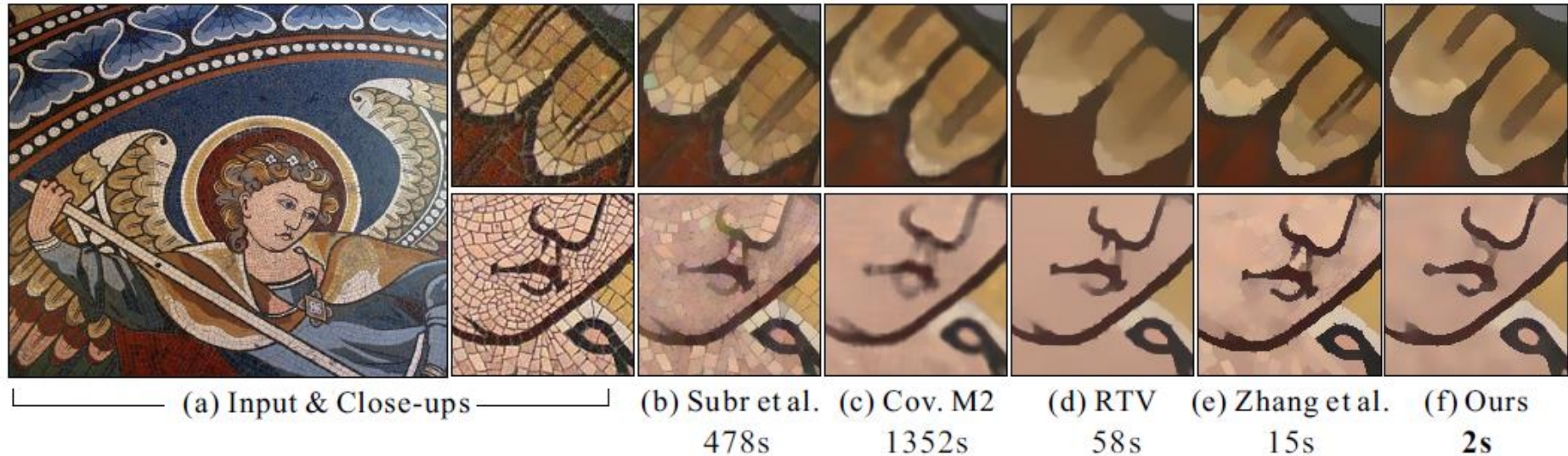


Fig. 12. Texture smoothing results and close-ups. (b)-(e) are results of [25], [14], [30], and [34] respectively. Parameters are (b) $k = 5$, (c) $\sigma = 0.3$, $k = 9$, (d) $\lambda = 0.015$, $\sigma = 5$, (e) $r = 3$, $\sigma_r = 0.3$, 10 iterations, (f) $\sigma_s = 5$, $\sigma_r = 0.1$.

Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Морфологическая фильтрация изображений

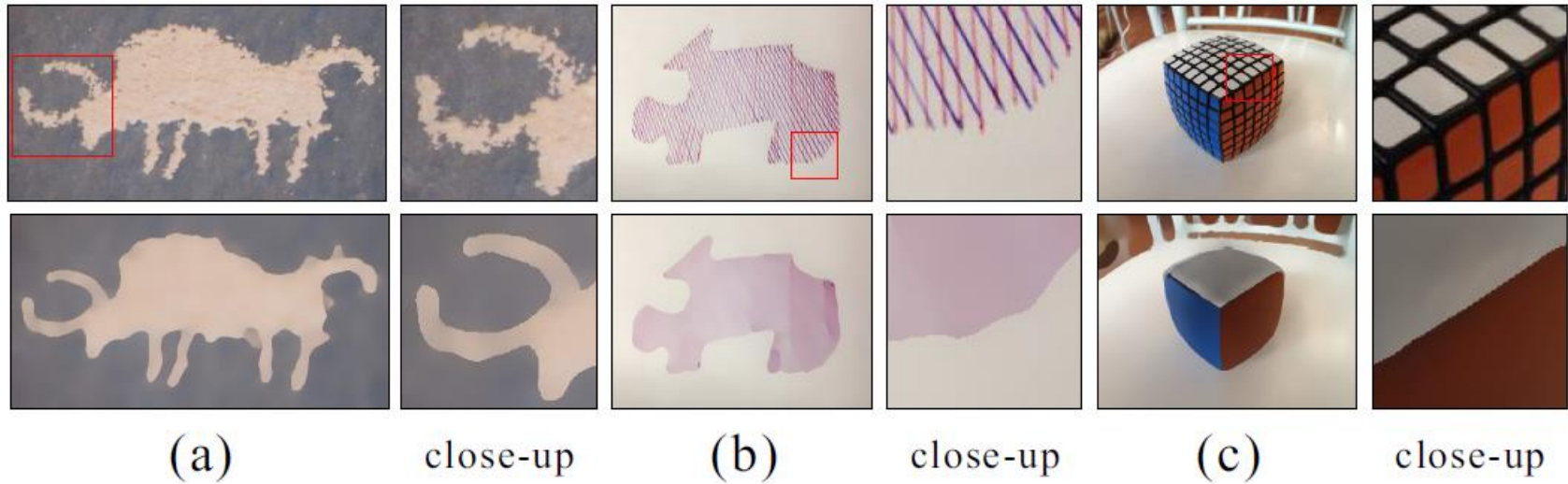
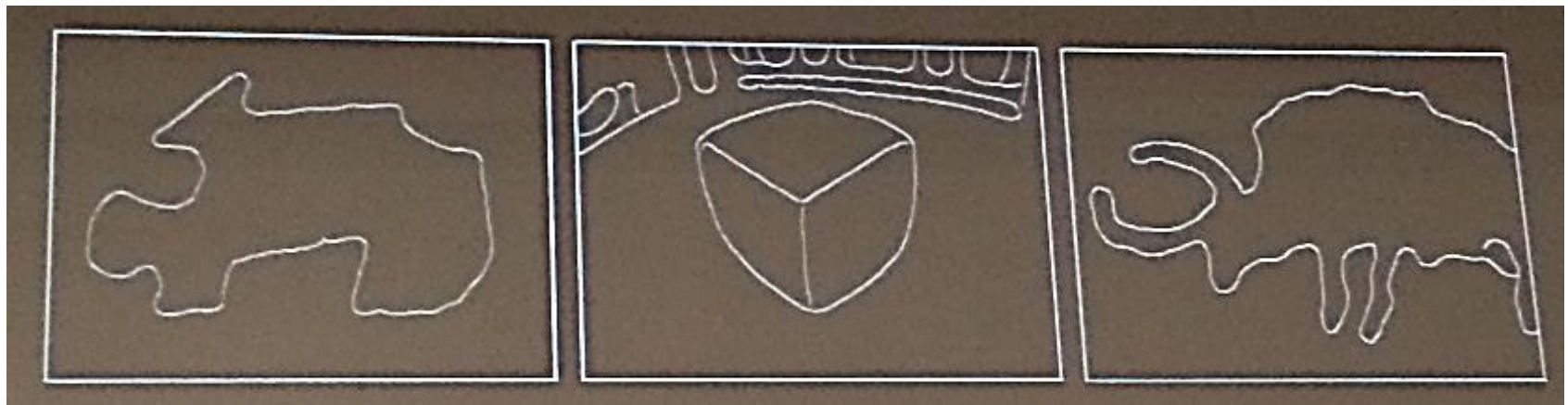


Fig. 13. Virtual contour restoration. Large contrast naturally forms region boundaries in human perception. Our filter can simulate this process.



Rolling Guidance Filter,

Qi Zhang, Xiaoyong Shen, Li Xu, Jiaya Jia

Shape Analysis

W26 Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA)

Organizers: Alex Bronstein
Umberto Castellani
Maks Ovsjanikov

1100 Anisotropic Laplace-Beltrami Operator for Shape Analysis, *Mathieu Andreux, Emanuele Rodolà, Mathieu Aubry, Daniel Cremers*

1150 A bioinformatics approach to 3D shape matching, *Manuele Bicego, Stefano Danese, Simone Melzi, Umberto Castellani*

1500 A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions, *Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, Stefano Berretti*

1600 A Novel Graph Embedding Framework for 3D Object Recognition, *Mario Manzo, Simone Pellino, Alfredo Petrosino, Alessandro Rozza*

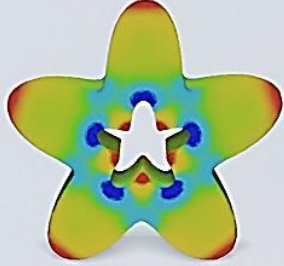
1625 Multiple Alignment of Spatiotemporal Deformable Objects for the Average-Organ Computation, *Atsushi Imiya, Shun Inagaki, Hayato Itoh*

1640 Refining Mitochondria Segmentation in EM Imagery with Active Surfaces, *Anne Jorstad, Pascal Fua*

Shape Analysis

Related work

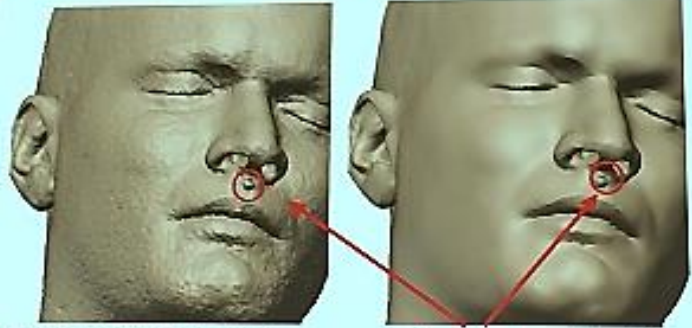
Isotropic spectral 3D shape analysis
↳ Intrinsic information



[Rustamov, SGP 07] [Sun et al, SGP 09] [Aubry et al, ICCV WS 11]
[Litman et al., C&G 11] [Rodolà et al., SGP 14]

Related work

Anisotropic Denoising / Regularization
↳ Curvature-aware



Anisotropic Denoising → Heeds local details
[Desbrun et al, SIGGRAPH 99] [Clarenz et al, VIS 00] [Tadiszyn et al, VIS 02]

Related work

High dimensional data analysis
by spectral clustering

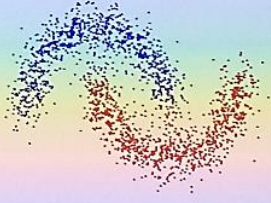
Anisotropy [Kim et al. ICCV 13]

↓

Reweighted graph Laplacian

↓

Transfer to shape analysis?



Limitations: non-linearity
ad hoc formulation

Goal

Extrinsic information + Intrinsic information

↓

Anisotropic Laplace-Beltrami operator

Keep Laplacian's nice mathematical properties !

Anisotropic Laplace-Beltrami Operator for Shape Analysis, Mathieu Andreux, Emanuele Rodolà, Mathieu Aubry, Daniel Cremers, ECCV'14, W26

Shape Analysis

Anisotropic Laplace-Beltrami Operators

Desired result: heat diffusion example

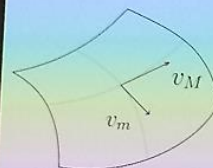


Isotropic

Anisotropic
High bending++

Anisotropic
Low bending++

Anisotropic Laplace-Beltrami Operators



Principal curvatures' directions

In practice:

$$D_\alpha = \begin{pmatrix} \frac{1}{1+\alpha|\kappa_M|} & 0 \\ 0 & \frac{1}{1+\alpha|\kappa_m|} \end{pmatrix} \begin{matrix} v_m \\ v_M \end{matrix}$$

Deviation from isotropy

$$\Delta_D = \text{div}(\underline{D} \nabla f)$$

Anisotropic tensor

D controls the direction/intensity of diffusion/wave propagation/...

Anisotropic Laplace-Beltrami Operators

Ours
 $\Delta_D = \text{div}(D \nabla f)$

$$D_\alpha = \begin{pmatrix} \frac{1}{1+\alpha|\kappa_M|} & 0 \\ 0 & \frac{1}{1+\alpha|\kappa_m|} \end{pmatrix}$$

Kim et al.'s
 $\Delta_{\tilde{D}} = \text{div}(\tilde{D}(\nabla f))$

$$\tilde{D} : \vec{v} \mapsto \frac{\|S\vec{v}\|}{\|\vec{v}\|} \vec{v}$$

$$S = \begin{pmatrix} \frac{1}{1+|\kappa_m|} & 0 \\ 0 & \frac{1}{1+|\kappa_M|} \end{pmatrix}$$

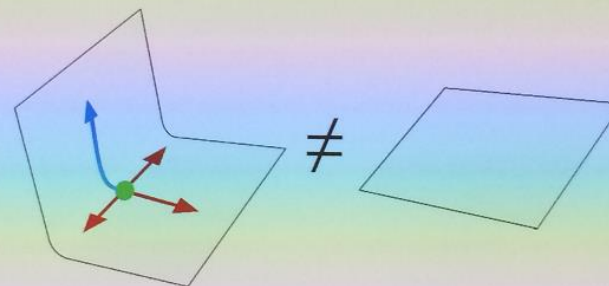
☺ Linear
Self-adjoint
Negative semi-definite
Eigendecomposition

Continuous formulation

Non-linear

Properties


Loss of isometry-invariance



We can actually differentiate these shapes!


Shape Analysis

Experiments




Datasets

TOSCA
(Michael)



[Bronstein et al. 08]

SHREC 10: 3 shapes, 9 deformations (incl. noise)

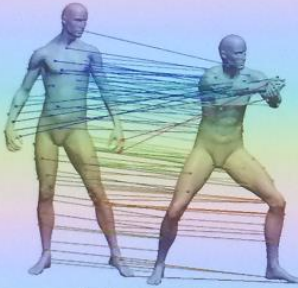


[Bronstein et al., EG 3DOR 10]

Matching

Goal

Matching points under near isometries



Procedure

Sample (*null*)

↓

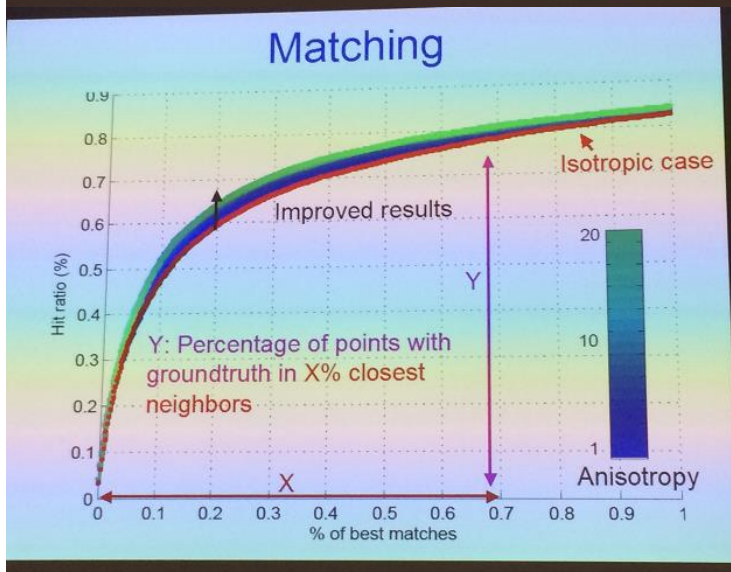
Descriptor on \neq poses

↓

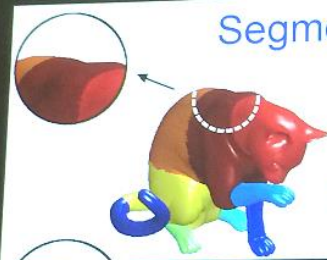
Order *null* points wrt descriptor similarity

↓

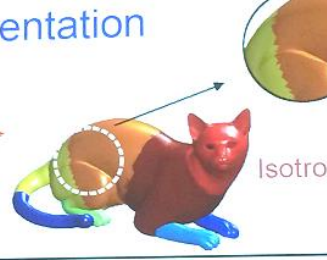
The sooner the groundtruth, the better!



Segmentation



Isotropic



Anisotropic

Consensus segmentation

[Rodolà et al. SGP'14]

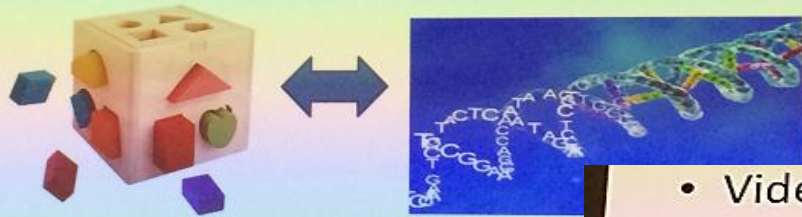
Anisotropic Laplace-Beltrami Operator for Shape Analysis, Mathieu Andreux, Emanuele Rodolà, Mathieu Aubry, Daniel Cremers, ECCV'14, W26

Shape Analysis

Overall aim

«can we exploit well-established bioinformatics tools to solve computer vision and pattern recognition problems?»

- We explore bioinformatics solutions to face 3D shape matching.



General idea

- **Idea:** encode the problem of 3D shape matching in biological terms and solve it as a sequence alignment problem



The 3D shape is transformed into a biological sequence

Two main sub-problems:

1. How to find an ordering from a 3D mesh
2. How to observe discrete symbols from a shape

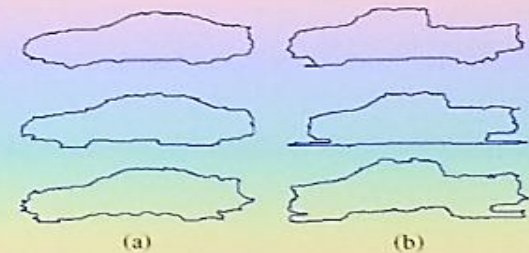
Similar approaches

• Video Genome



Bronstein, A.M., Bronstein, M.M., Kimmel, R.: The video genome. CoRR abs/1003.5320 (2010)

• 2D shape matching



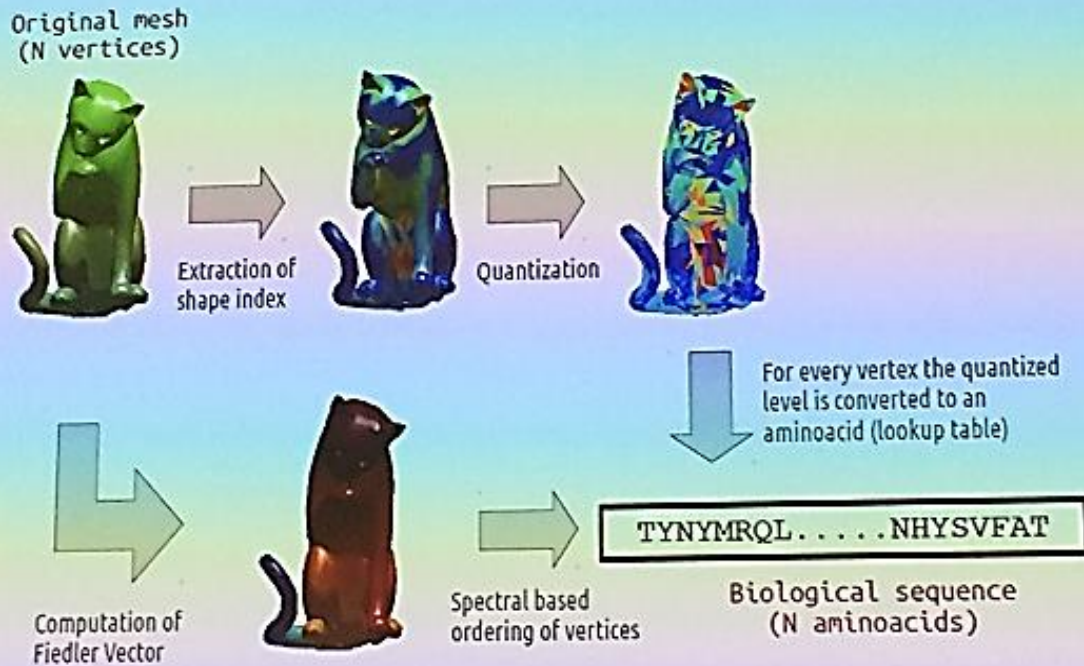
Bicego, M., Lovato, P.: 2d shape recognition using biological sequence alignment tools. In: ICPR. pp. 1359–1362

A bioinformatics approach to 3D shape matching,

Manuele Bicego, Stefano Danese, Simone Melzi, Umberto Castellani, ECCV'14, W26

Shape Analysis

Proposed pipeline



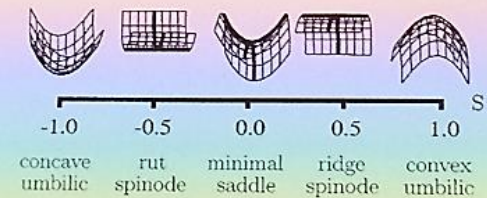
Local feature extraction

- **Shape Index:** represents the degree of concavity of a surface patch:

$$s = -\frac{2}{\pi} \arctan \left(\frac{k_1 + k_2}{k_1 - k_2} \right) \quad k_1 > k_2$$

k_1 and k_2 are the principal curvatures.

- According to the Shape index value the surface can be classified as:



Feature quantization

- We define two very simple schemes:
 - ‘DNA-mapping’: shape index values are mapped to the 3 intervals by leading to 3 symbols.
 - ‘Protein mapping’: shape index values are mapped as the aminoacid alphabet composed by 20 symbols.



A bioinformatics approach to 3D shape matching,

Manuele Bicego, Stefano Danese, Simone Melzi, Umberto Castellani, ECCV'14, W26

Shape Analysis

Spectral based ordering

- As proposed for streaming mesh, or mesh partitioning we used the ordering provided by the second eigenvector of Laplace operator, i.e., the so called *Fiedler Vector*
- Fiedler vector provides an heuristic solution to the *minimum linear arrangement* (MLA) problem.

$$v_2 = \begin{cases} \operatorname{argmin}_{u \in \mathbb{R}} \sum_{i,j=1}^N w_{ij}(u_i - u_j)^2 \\ \text{s.t. } u'e = 0, u'u = 1 \end{cases}$$

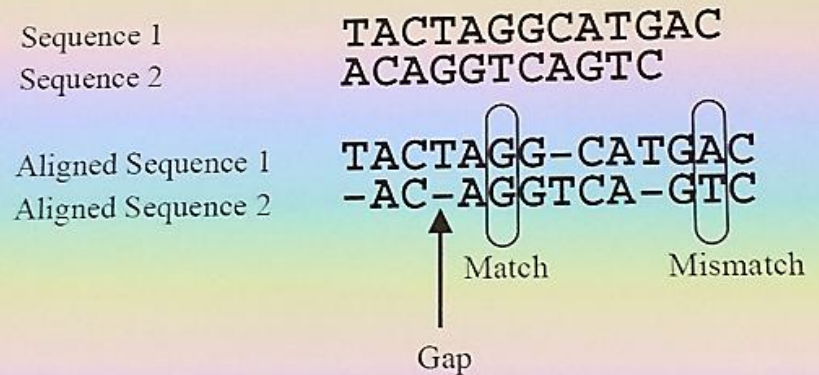
Spectral based ordering

Fiedler Vector



Sequence alignment

- **Needleman-Wunsh** (NW) algorithm: is a dynamic programming method for finding the best *global* alignment between two sequences



- We evaluate both **Smith-Waterman** (SW) and **Needleman-Wunch** (NW) algorithms
- We compare our approach with:
 - **Shape DNA**,
 - **Dynamic Time Warping (DTW)** distance between Shape index sequeces ordered by Fiedler Vector,
 - **Histogram of Shape Index**.
- We perform shape classification using a **Nearest Neighbour** approach.

A bioinformatics approach to 3D shape matching,

Manuele Bicego, Stefano Danese, Simone Melzi, Umberto Castellani, ECCV'14, W26

Shape Analysis

Results

Method	AA	NT
NW (Basic)	0.0000	0.0350
SW (Basic)	0.0629	0.1189
NW (Advanced)	0.0000	0.1888
SW (Advanced)	0.0629	0.2308

Performance evaluation

Method	Error LOO
Shape DNA	0.0070
Shape Index Hist (100 bin)	0.0839
DTW	0.0420
Proposed approach (best)	0.0000

Comparison with other methods



Results



Point to point matching

Tosca Dataset: is composed by 10 classes of non-rigid objects: cat, centaur, man1, dog, gorilla, man2, horse, lioness, seahorse, and woman

A bioinformatics approach to 3D shape matching,

Manuele Bicego, Stefano Danese, Simone Melzi, Umberto Castellani, ECCV'14, W26

Shape Analysis

Results

Isometry transformation

Method	AA	NT
NW (Basic)	0.0408	0.1224
SW (Basic)	0.0612	0.1429
NW (Advanced)	0.0000	0.1020
SW (Advanced)	0.0000	0.1020

Performance evaluation

Isometry-topology transformation

Method	AA	NT
NW (Basic)	0.0000	0.1837
SW (Basic)	0.0000	0.2245
NW (Advanced)	0.0000	0.1224
SW (Advanced)	0.0000	0.1224

Performance evaluation

Method	Error LOO
Shape DNA	0.1020
Shape Index Hist (20 bin)	0.1837
DTW	0.0408
Proposed approach (best)	0.0000

Comparison with other methods

Method	Error LOO
Shape DNA	0.3469
Shape Index Hist (20 bin)	0.2041
DTW	0.1224
Proposed approach (best)	0.0000

Comparison with other methods

Results



Shape Google Dataset: is composed by 10 classes of non-rigid objects: dog, cat1, cat2, woman, man, dromedary, elephant, flamingo, horse, gouger. Each object appears with multiple modifications and transformations of the original shape.

A bioinformatics approach to 3D shape matching,

Manuele Bicego, Stefano Danese, Simone Melzi, Umberto Castellani, ECCV'14, W26

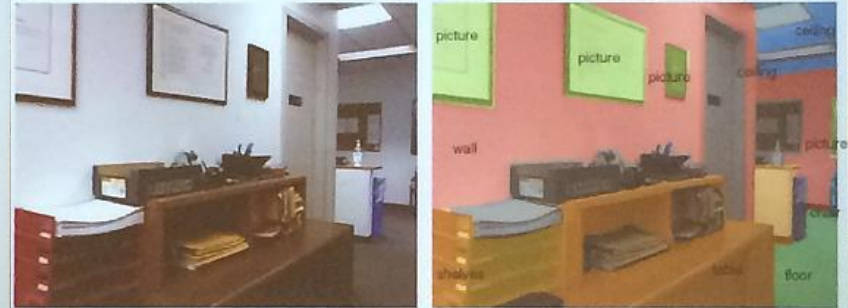
Shape Analysis, Energy-based, SPD

Goal: Segmentation, Recognition

Segment images into regions corresponding to projections of different objects in a real 3d scene and *recognize* the objects

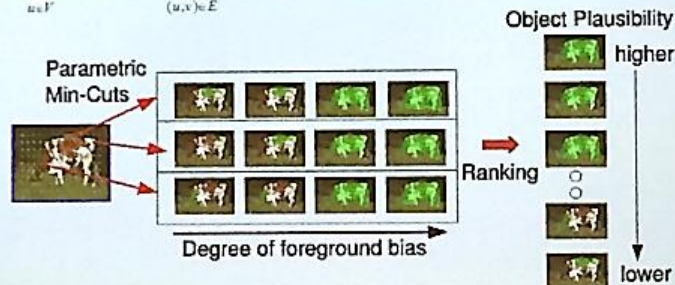


Scene Understanding



CPMC: Constrained Parametric Min-Cuts for Automatic Object Segmentation

$$E_{\lambda}(x) = \sum_{u \in F} D(x_u, \lambda) + \sum_{(u,v) \in E} V_{uv}(x_u, x_v) \rightarrow \min_{x, \lambda} E_{\lambda}(x)$$



Solve for all **breakpoints** (x, λ) using parametric max flow

Carreira and Sminchisescu, ICCV09, CVPR 10, PAMI 11

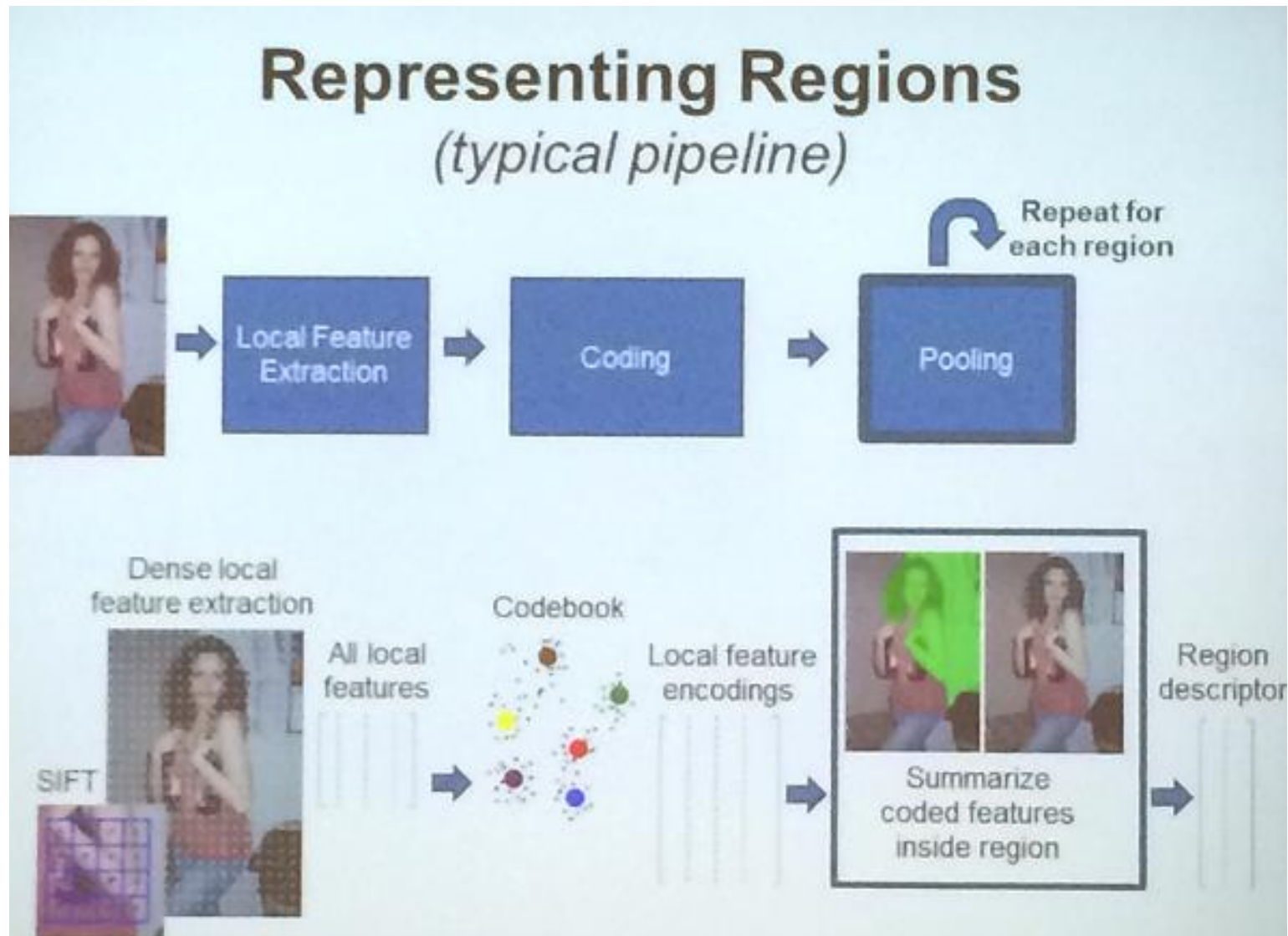
Ideally, we would want a framework that `uniformly` accommodates color, depth and video analysis

Problems

- Region generation
 - Systematic, combinatorial
 - Boundaries from RGB, depth, motion
- Region selection, pool compression
 - Object-like=class-independent=objectness
 - Maximum marginal diversification
- Region description
 - Second-order methods
- Complete scene recognition by composition
 - Re-combination, re-segmentation
 - Sequential vs. simultaneous

A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions, Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, Stefano Berretti, ECCV'14, W26

Shape Analysis, Energy-based, SPD

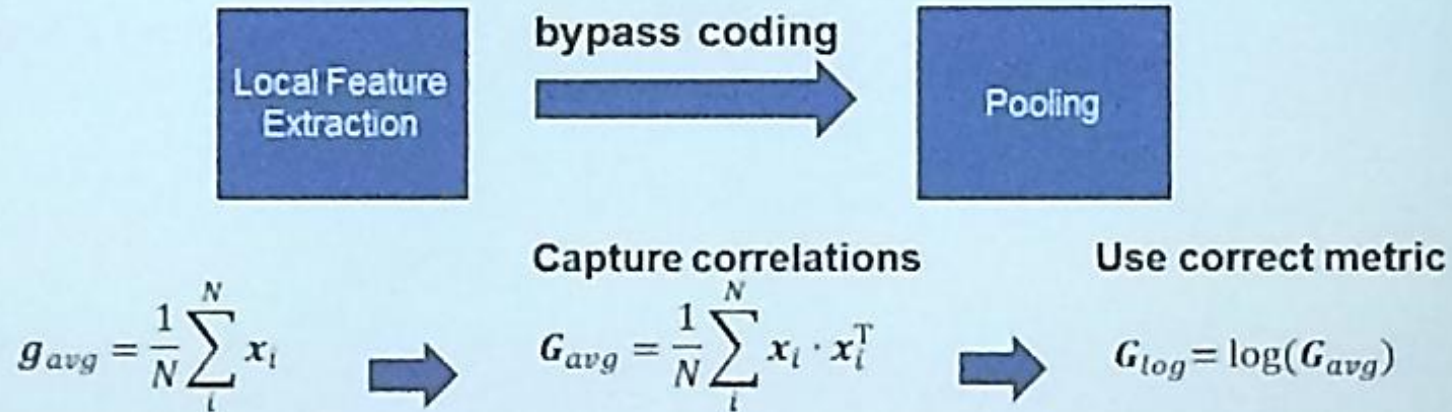


A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions, Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, Stefano Berretti, ECCV'14, W26

Shape Analysis, Energy-based, SPD

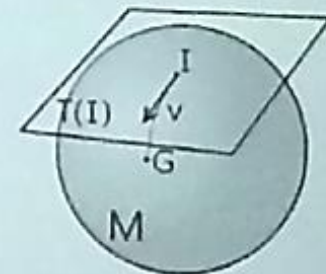
Second Order Pooling (O2P)

Can we pursue higher-order statistics for pooling ?



Using **Log-Euclidean metric** we can directly embed entire manifold of SPD matrices

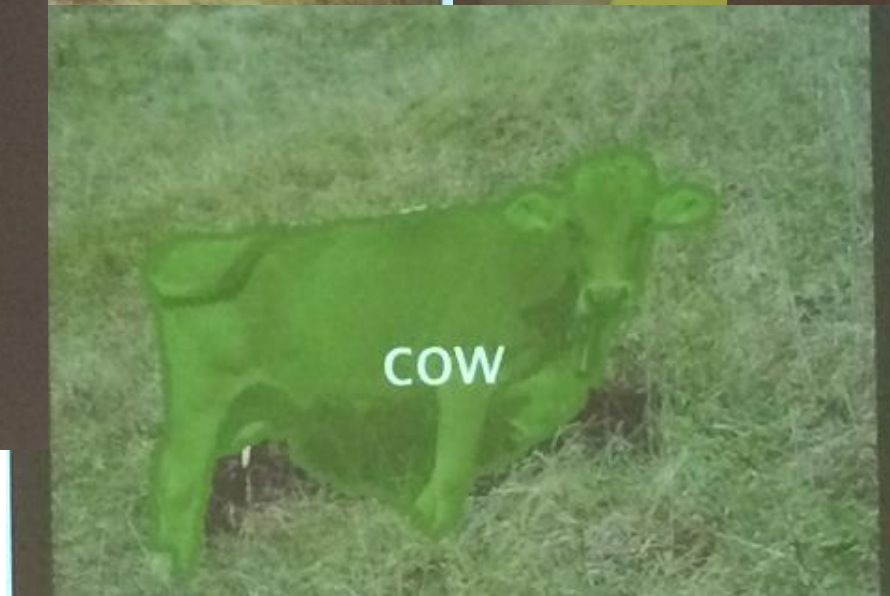
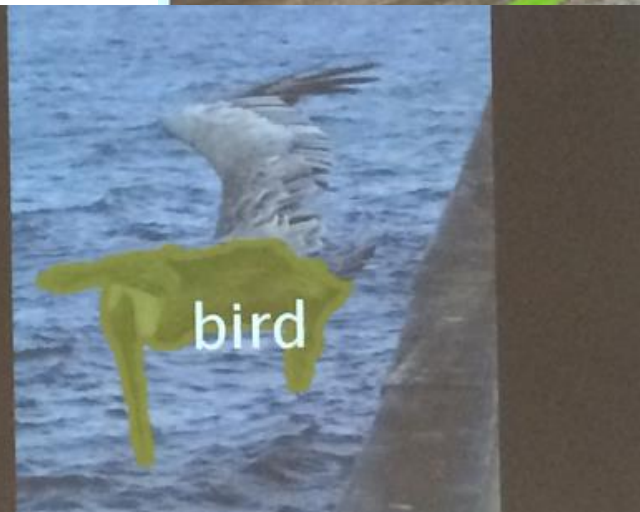
Dimensionality = (local descriptor size)²



Carreira, Caseiro, Batista, Sminchisescu, ECCV12

A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions, Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, Stefano Berretti, ECCV'14, W26

Shape Analysis, Energy-based, SPD



A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions, Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, Stefano Berretti, ECCV'14, W26

SPD, Manifolds

CCA IN EUCLIDEAN SPACE

Pearson correlation for $x \in \mathbf{R}$ and $y \in \mathbf{R}$

$$\rho_{x,y} = \frac{\text{COV}(x,y)}{\sigma_x \sigma_y} = \frac{\mathbb{E}[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^N (y_i - \mu_y)^2}} \quad (1)$$

Canonical Correlation for $\mathbf{x} \in \mathbf{R}^m$ and $\mathbf{y} \in \mathbf{R}^n$

$$\max_{\mathbf{w}_x, \mathbf{w}_y} \text{corr}(\pi_{\mathbf{w}_x}(\mathbf{x}), \pi_{\mathbf{w}_y}(\mathbf{y})) = \max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\sum_{i=1}^N \mathbf{w}_x^T (\mathbf{x}_i - \mu_x) \mathbf{w}_y^T (\mathbf{y}_i - \mu_y)}{\sqrt{\sum_{i=1}^N (\mathbf{w}_x^T (\mathbf{x}_i - \mu_x))^2} \sqrt{\sum_{i=1}^N (\mathbf{w}_y^T (\mathbf{y}_i - \mu_y))^2}} \quad (2)$$

where projection coefficient $\pi_{\mathbf{w}_x}(\mathbf{x}) := \arg \min_{t \in \mathbf{R}} d(t\mathbf{w}_x + \mu_x, \mathbf{x})^2$.

CCA ON MANIFOLDS: BASIC OPERATIONS

Operation	Subtraction	Addition	Distance	Mean	Covariance
Euclidean	$\vec{x_i x_j} = x_j - x_i$	$x_i + \vec{x_j x_k}$	$\ \vec{x_i x_j}\ $	$\sum_{i=1}^n \vec{x_i x_i} = 0$	$\mathbb{E}[(x_i - \bar{x})(x_i - \bar{x})^T]$
Riemannian	$\vec{x_i x_j} = \text{Log}(x_i, x_j)$	$\text{Exp}(x_i, \vec{x_j x_k})$	$\ \text{Log}(x_i, x_j)\ _{x_i}$	$\sum_{i=1}^n \text{Log}(\bar{x}, x_i) = 0$	$\mathbb{E}[\text{Log}(\bar{x}, x_i) \text{Log}(\bar{x}, x_i)^T]$

$$\pi_{\mathbf{w}_x}(\mathbf{x}) := \arg \min_{t \in \mathbf{R}} d(\text{Exp}(\mu_x, t\mathbf{w}_x), \mathbf{x})^2 \quad (3)$$

Canonical Correlation Analysis on Riemannian Manifolds and Its Applications,
 Hyunwoo J. Kim, Nagesh Adluru, Barbara B. Bendlin, Sterling C. Johnson,
 Baba C. Vemuri, and Vikas Singh, ECCV'14

SPD, Manifolds

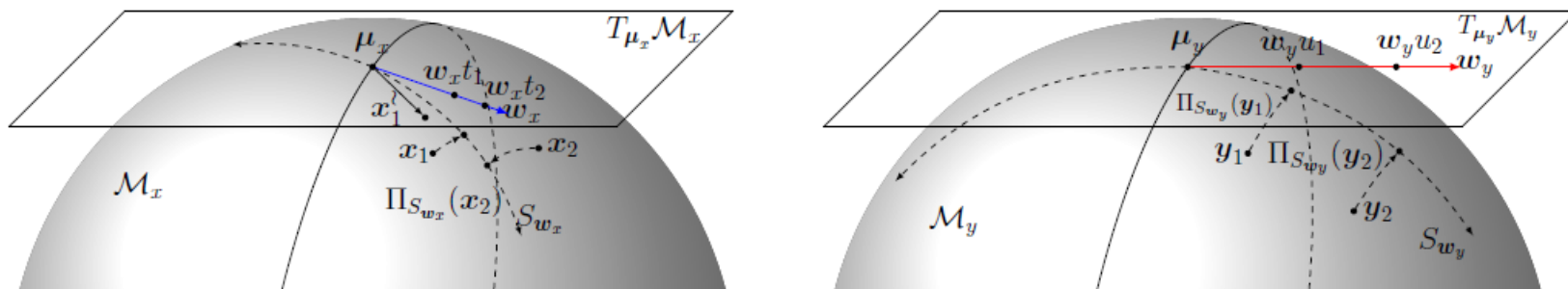


Fig. 1. CCA on Riemannian manifolds. CCA searches geodesic submanifolds (subspaces), S_{w_x} and S_{w_y} at the Karcher mean of data on each manifold. Correlation between projected points $\{\Pi_{S_{w_x}}(\mathbf{x}_i)\}_{i=1}^N$ and $\{\Pi_{S_{w_y}}(\mathbf{y}_i)\}_{i=1}^N$ is equivalent to the correlation between *projection coefficients* $\{t_i\}_{i=1}^N$ and $\{u_i\}_{i=1}^N$. Although \mathbf{x} and \mathbf{y} belong to the same manifold we show them in different plots for ease of explanation.

$$\rho_{\mathbf{x}, \mathbf{y}} = \max_{w_x, w_y, t, u} \frac{\sum_{i=1}^N (t_i - \bar{t})(u_i - \bar{u})}{\sqrt{\sum_{i=1}^N (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^N (u_i - \bar{u})^2}} \quad (10)$$

$$s.t. \quad t_i = \arg \min_{t_i \in (-\epsilon, \epsilon)} \|\text{Log}(\text{Exp}(\mu_x, t_i w_x), \mathbf{x}_i)\|^2, \forall i \in \{1, \dots, N\}$$

$$u_i = \arg \min_{u_i \in (-\epsilon, \epsilon)} \|\text{Log}(\text{Exp}(\mu_y, u_i w_y), \mathbf{y}_i)\|^2, \forall i \in \{1, \dots, N\}$$

SPD, Manifolds

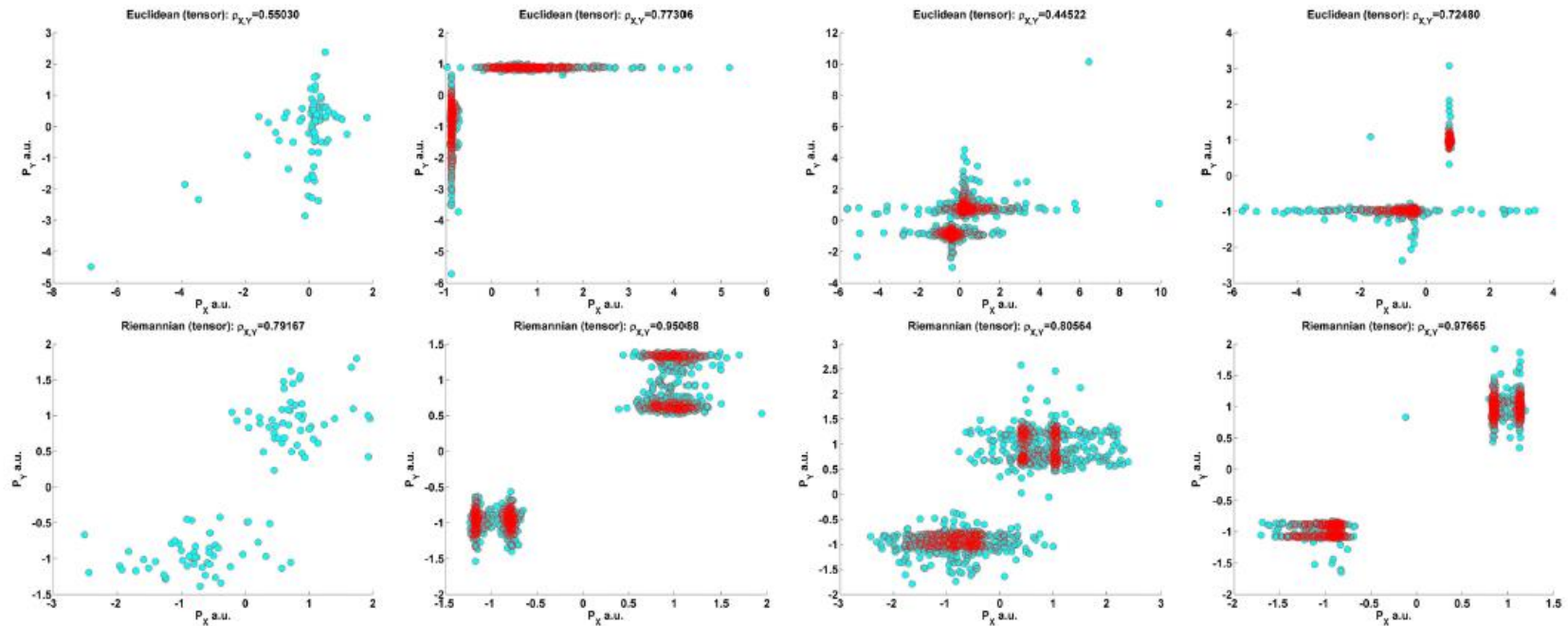
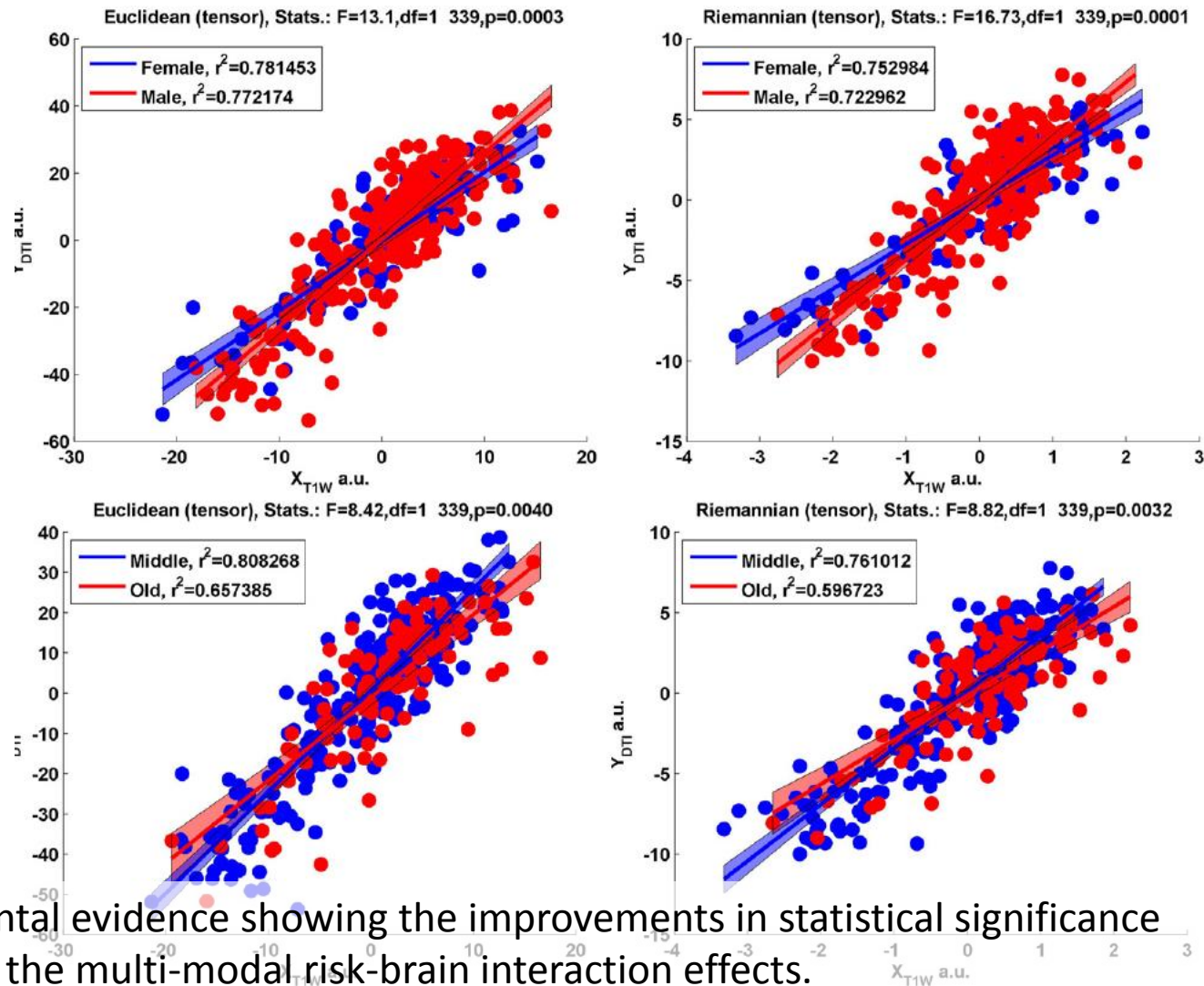


Fig. 2. Synthetic experiments showing the benefits of Riemannian CCA. The top row shows the projected data using the Euclidean CCA and the bottom using Riemannian CCA. P_X and P_Y denote the projected axes. Each column represents a synthetic experiment with a specific set of $\{\mu_{x_j}, \epsilon_{x_j}; \mu_{y_j}, \epsilon_{y_j}\}$. The first column presents results with 100 samples while the three columns on the right show with 1000 samples. The improvements in the correlation coefficients $\rho_{x,y}$ can be clearly seen from the corresponding titles.

Canonical Correlation Analysis on Riemannian Manifolds and Its Applications,
Hyunwoo J. Kim, Nagesh Adluru, Barbara B. Bendlin, Sterling C. Johnson,
Baba C. Vemuri, and Vikas Singh, ECCV'14

SPD, Manifolds



Experimental evidence showing the improvements in statistical significance of finding the multi-modal risk-brain interaction effects.

Canonical Correlation Analysis on Riemannian Manifolds and Its Applications,
Hyunwoo J. Kim, Nagesh Adluru, Barbara B. Bendlin, Sterling C. Johnson,
Baba C. Vemuri, and Vikas Singh, ECCV'14

SPD, Manifolds

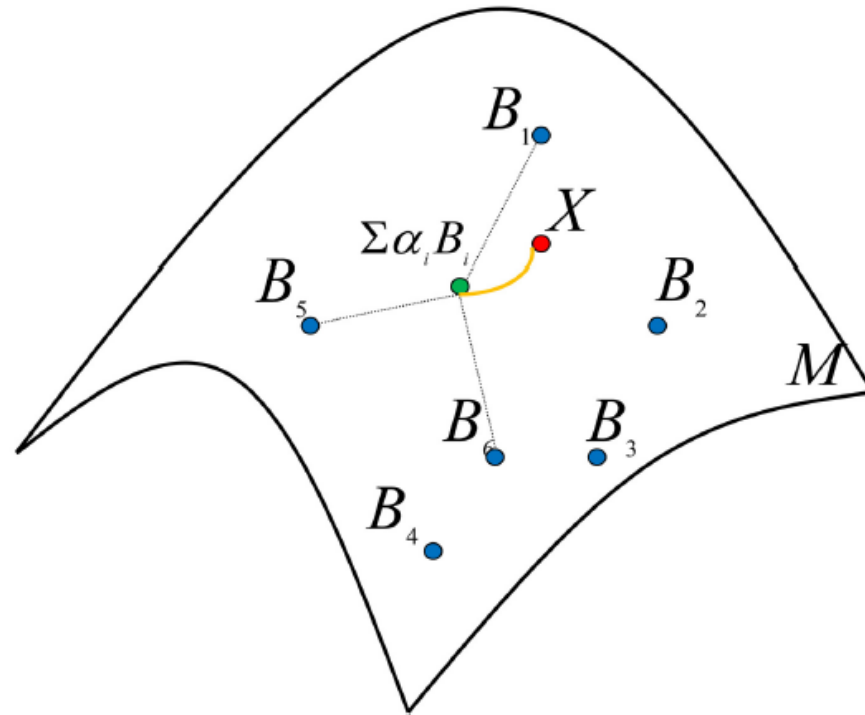
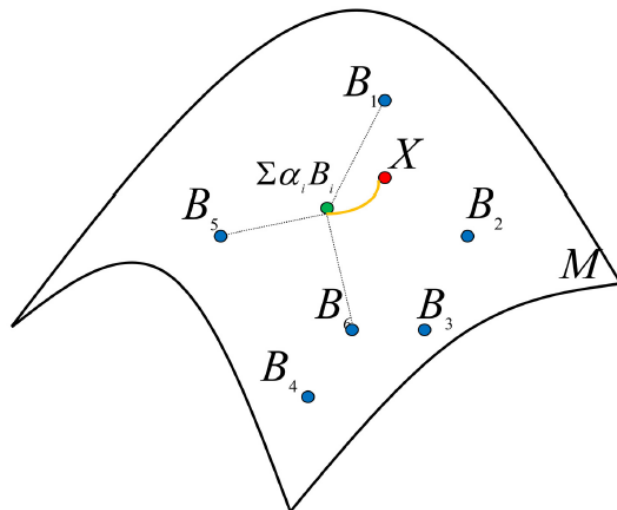


Fig. 1. A schematic illustration of our sparse coding objective formulation. For the SPD manifold M and given SPD basis matrices B_i on the manifold, our objective seeks a non-negative sparse linear combination $\sum_i \alpha_i B_i$ of the B_i 's that is closest (in a geodesic sense) to the given input SPD matrix X .

SPD, Manifolds



Model. Let \mathcal{B} be a dictionary with n atoms B_1, B_2, \dots, B_n , where each $B_i \in \mathcal{S}_+^d$. Let $X \in \mathcal{S}_+^d$ be an input matrix that must be sparse coded. Our basic sparse coding objective is to solve

$$\begin{aligned} \min_{\alpha \geq 0} \quad \phi(\alpha) &:= \frac{1}{2} d_{\mathcal{R}}^2 \left(\sum_{i=1}^n \alpha_i B_i, X \right) + \text{Sp}(\alpha) \\ &= \frac{1}{2} \left\| \text{Log} \sum_{i=1}^n \alpha_i X^{-\frac{1}{2}} B_i X^{-\frac{1}{2}} \right\|_{\text{F}}^2 + \text{Sp}(\alpha), \end{aligned}$$

where α_i is the i -th component of α , and Sp is a sparsity inducing function.

Riemannian Sparse Coding for Positive Definite Matrices,
Anoop Cherian and Suvrit Sra, ECCV'14

SPD, Manifolds

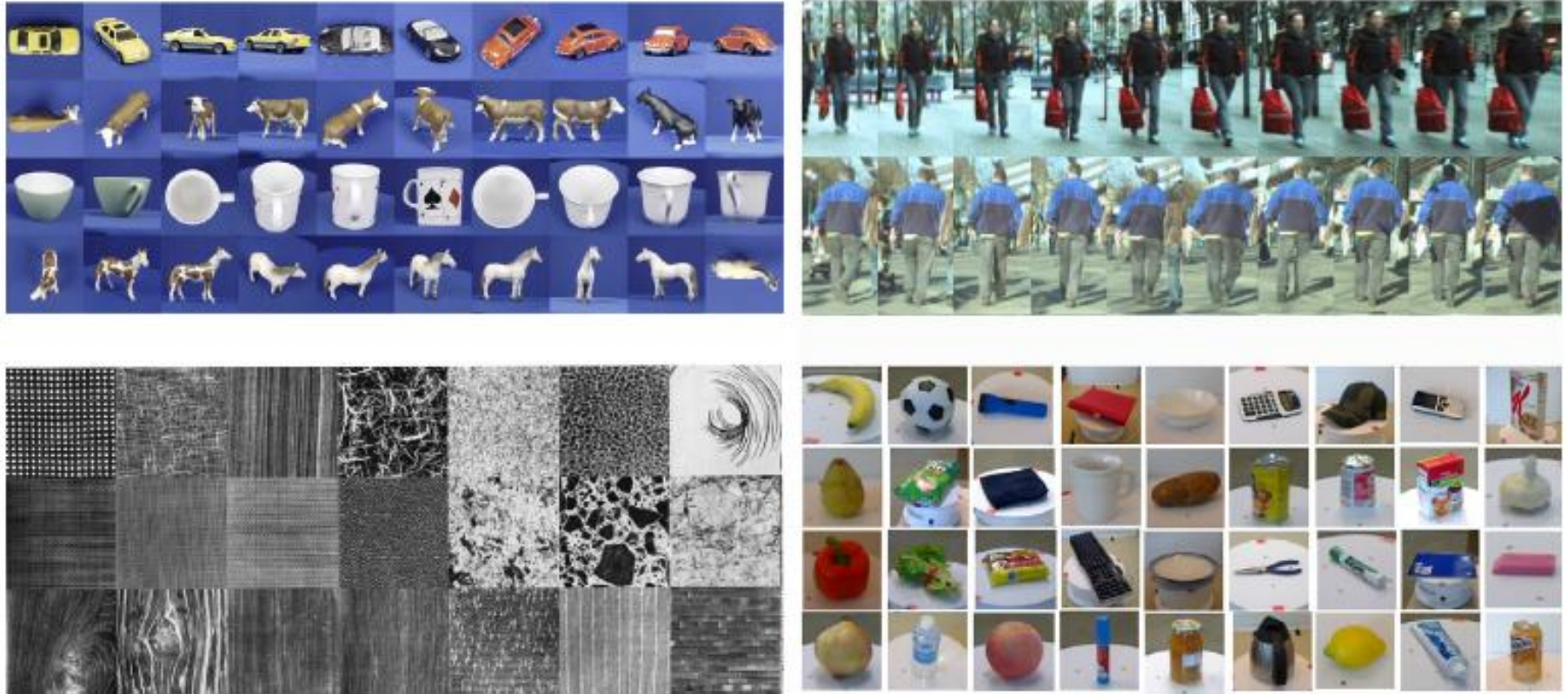


Fig. 3. Montage of sample images from the four datasets used in our experiments. Top-left are samples from the ETH80 object dataset, bottom-left are the Brodatz textures, top-right are the samples from ETHZ people dataset, and images from the RGB-D object recognition dataset are shown on bottom right.

SPD, Manifolds

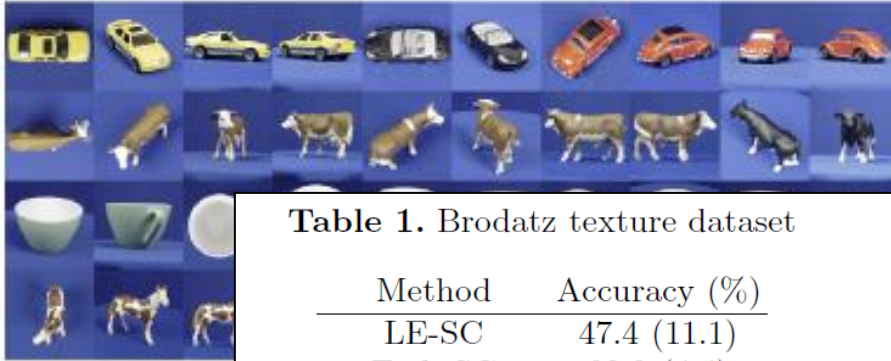


Table 1. Brodatz texture dataset

Method	Accuracy (%)
LE-SC	47.4 (11.1)
Frob-SC	32.3 (4.4)
K-Stein-SC	39.2 (0.79)
K-LE-SC	47.9 (0.46)
TSC	35.6 (7.1)
GDL	43.7 (6.3)
Riem-SC(ours)	53.9 (3.4)



Table 2. ETH80 object recognition

Method	Accuracy (%)
LE-SC	68.9 (3.3)
Frob-SC	67.3 (1.4)
K-Stein-SC	81.6 (2.1)
K-LE-SC	76.6 (0.4)
TSC	37.1 (3.9)
GDL	65.8 (3.1)
Riem-SC(ours)	77.9 (1.9)



Table 3. ETHZ Person Re-identification

Method	Accuracy (%)
LE-SC	78.5 (2.5)
Frob-SC	83.7 (0.2)
K-Stein-SC	88.3 (0.4)
K-LE-SC	87.8 (0.8)
TSC	67.7 (1.2)
GDL	30.5 (1.7)
Riem-SC(ours)	90.1 (0.9)

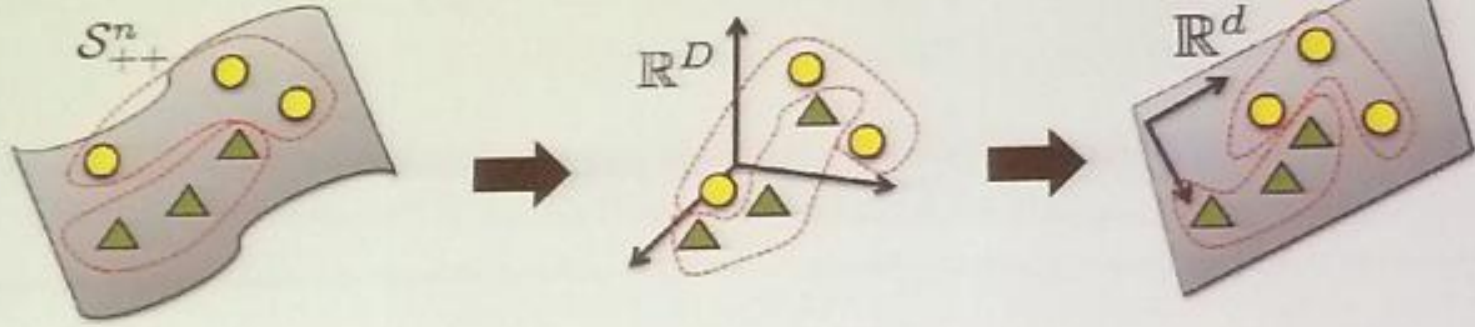
Table 4. RGB-D Object Recognition

Method	Accuracy (%)
LE-SC	86.1 (1.0)
Frob-SC	80.3 (1.1)
K-Stein-SC	75.6 (1.1)
K-LE-SC	83.5 (0.2)
TSC	72.8 (2.1)
GDL	61.9 (0.4)
Riem-SC(ours)	84.0 (0.6)



SPD, Manifolds

Dimensionality reduction methods on the manifolds



Principal Geodesic Analysis
(Fletcher et al. TMI'04):

- Generalization of PCA to Riemannian manifolds
- Unsupervised method
- Flattens the manifold at the Karcher mean

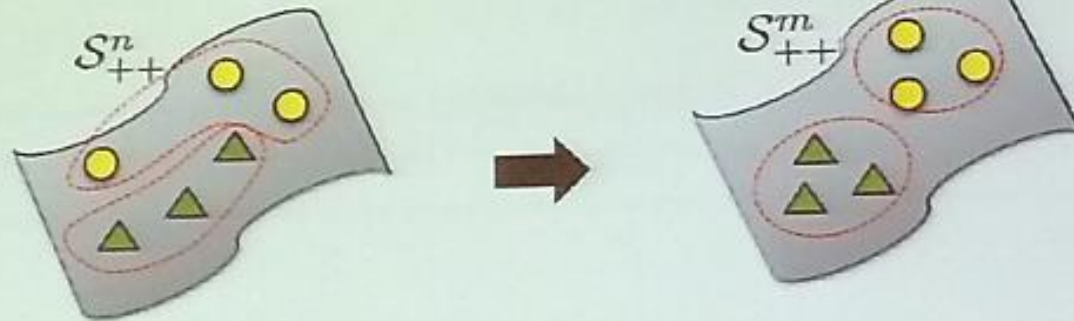
Covariance Discriminative Learning
(Wang et al. CVPR'12):

- Designed for SPD matrices
- Finds a discriminative Euclidean subspace
- Relies on principal matrix logarithm (tangent space)

From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices,
Mehrtash Harandi, Mathieu Salzmann, Richard Hartley, ECCV'14

SPD, Manifolds

Proposed dimensionality reduction technique



We learn a representation that

- is low-dimensional and discriminative
- still benefits from useful properties of SPD matrices
- can be used in conjunction with existing manifold-based techniques

From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices,
Mehrtash Harandi, Mathieu Salzmann, Richard Hartley, ECCV'14

SPD, Manifolds

Proposed dimensionality reduction technique

Geometry-Aware Dimensionality Reduction

We now describe our approach to learning an embedding of high-dimensional SPD matrices to a more discriminative, low-dimensional SPD manifold. More specifically, given a matrix $\mathbf{X} \in \mathcal{S}_{++}^n$, we seek to learn the parameters $\mathbf{W} \in \mathbb{R}^{n \times m}$, $m < n$, of a generic mapping $f : \mathcal{S}_{++}^n \times \mathbb{R}^{n \times m} \rightarrow \mathcal{S}_{++}^m$, which we define as

$$f(\mathbf{X}, \mathbf{W}) = \mathbf{W}^T \mathbf{X} \mathbf{W}. \quad (4)$$

Clearly, if $\mathcal{S}_{++}^n \ni \mathbf{X} \succ 0$ and \mathbf{W} has full rank, $\mathcal{S}_{++}^m \ni \mathbf{W}^T \mathbf{X} \mathbf{W} \succ 0$.

Given a set of SPD matrices $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_p\}$, where each matrix $\mathbf{X}_i \in \mathcal{S}_{++}^n$, our goal is to find a transformation \mathbf{W} such that the resulting low-dimensional SPD manifold preserves some interesting structure of the original data. Here, we encode this structure via an undirected graph defined by a real symmetric affinity matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$. The element \mathbf{A}_{ij} of this matrix measures some notion of affinity between matrices \mathbf{X}_i and \mathbf{X}_j , and may be negative. We will discuss the affinity matrix in more details in Section 4.2.

From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices,
Mehrtash Harandi, Mathieu Salzmann, Richard Hartley, ECCV'14

SPD, Manifolds

Proposed dimensionality reduction technique

To avoid degeneracies and ensure that the resulting embedding forms a valid SPD manifold, *i.e.*, $\mathbf{W}^T \mathbf{X} \mathbf{W} \succ 0$, $\forall \mathbf{X} \in \mathcal{S}_{++}^n$, we need \mathbf{W} to have full rank. Here, we enforce this requirement by imposing orthonormality constraints on \mathbf{W} , *i.e.*, $\mathbf{W}^T \mathbf{W} = \mathbf{I}_m$. Note that, with either the AIRM or the Stein divergence, this entails no loss of generality. Indeed, any full rank matrix $\tilde{\mathbf{W}}$ can be expressed as $\mathbf{M} \mathbf{W}$, with \mathbf{W} an orthonormal matrix and $\mathbf{M} \in \text{GL}(n)$. The affine invariance property of the AIRM and of the Stein metric therefore guarantees that

$$\mathcal{J}_{ij}(\tilde{\mathbf{W}}; \mathbf{X}_i, \mathbf{X}_j) = \mathcal{J}_{ij}(\mathbf{M} \mathbf{W}; \mathbf{X}_i, \mathbf{X}_j) = \mathcal{J}_{ij}(\mathbf{W}; \mathbf{X}_i, \mathbf{X}_j) .$$

Finally, learning can be expressed as the minimization problem

$$\mathbf{W}^* = \arg \min_{\mathbf{W} \in \mathbb{R}^{n \times m}} \sum_{i,j} \mathbf{A}_{ij} \delta^2 \left(\mathbf{W}^T \mathbf{X}_i \mathbf{W}, \mathbf{W}^T \mathbf{X}_j \mathbf{W} \right) \quad \text{s.t. } \mathbf{W}^T \mathbf{W} = \mathbf{I}_m$$

SPD, Manifolds

Testing relative to:

NN-Stein: Stein metric-based Nearest Neighbor classifier.

NN-AIRM: AIRM-based Nearest Neighbor classifier.

NN-Stein-ML: Stein metric-based Nearest Neighbor classifier on the low-dimensional SPD manifold obtained with our approach.

NN-AIRM-ML: AIRM-based Nearest Neighbor classifier on the low-dimensional SPD manifold obtained with our approach.

RSR: Riemannian Sparse Representation [6].

RSR-ML: Riemannian Sparse Representation on the low-dimensional SPD manifold obtained with our approach.

Results:

Table 1. Mean recognition accuracies with standard deviations for the UIUC material dataset [12]

Method	Accuracy
SD [12]	43.5% \pm N/A
CDL [22]	52.3% \pm 4.3
NN-Stein	35.8% \pm 2.6
NN-Stein-ML	58.1% \pm 2.8
NN-AIRM	35.6% \pm 2.6
NN-AIRM-ML	58.3% \pm 2.3
RSR [6]	52.8% \pm 2.1
RSR-ML	66.6% \pm 3.1

Material Categorization



Fig. 2. Samples from the UIUC material dataset

From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices,
Mehrtash Harandi, Mathieu Salzmann, Richard Hartley, ECCV'14

SPD, Manifolds

Testing relative to:

NN-Stein: Stein metric-based Nearest Neighbor classifier.

NN-AIRM: AIRM-based Nearest Neighbor classifier.

NN-Stein-ML: Stein metric-based Nearest Neighbor classifier on the low-dimensional SPD manifold obtained with our approach.

NN-AIRM-ML: AIRM-based Nearest Neighbor classifier on the low-dimensional SPD manifold obtained with our approach.

RSR: Riemannian Sparse Representation [6].

RSR-ML: Riemannian Sparse Representation on the low-dimensional SPD manifold obtained with our approach.

Results:

Table 2. Recognition accuracies for the HDM05-MOCAP dataset [14]

Method	Accuracy
CDL [22]	79.8%
NN-Stein	61.7%
NN-Stein-ML	68.6%
NN-AIRM	62.8%
NN-AIRM-ML	67.6%
RSR [6]	76.1%
RSR-ML	81.9%

Action Recognition from Motion Capture Data

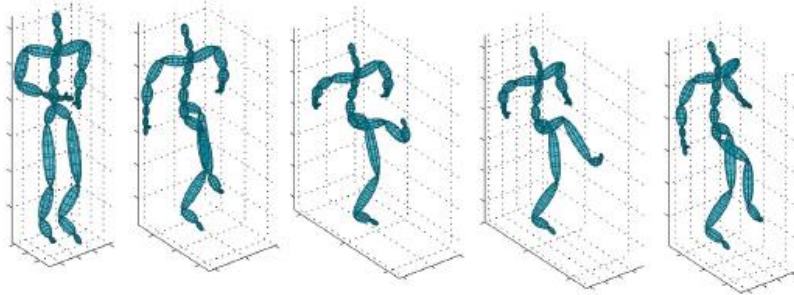


Fig. 3. Kicking action from the HDM05 motion capture sequences database [14]

From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices,
Mehrtash Harandi, Mathieu Salzmann, Richard Hartley, ECCV'14

SPD, Manifolds

Results:

Table 3. Recognition accuracies for the FERET face dataset [17]

Method	bc	bd	be	bf	bg	bh	average acc.
SRC [23]	9.5%	37.5%	77.0%	88.0%	48.5%	11.0%	45.3% \pm 3.3
GSRC [25]	35.5%	77.0%	93.5%	97.0%	79.0%	38.0%	70.0% \pm 2.7
CDL [22]	35.0%	87.5%	99.5%	100.0%	91.0%	34.5%	74.6% \pm 3.1
NN-Stein	29.0%	75.5%	94.5%	98.0%	83.5%	34.5%	69.2% \pm 3.0
NN-Stein-ML	40.5%	88.5%	97.0%	99.0%	91.5%	44.5%	76.8% \pm 2.7
NN-AIRM	28.5%	72.5%	93.0%	97.5%	83.0%	35.0%	68.3% \pm 3.0
NN-AIRM-ML	39.0%	84.0%	96.0%	99.0%	90.5%	45.5%	75.7% \pm 2.6
RSR [6]	36.5%	79.5%	96.5%	97.5%	86.0%	41.5%	72.9% \pm 2.7
RSR-ML	49.0%	90.5%	98.5%	100%	93.5%	50.5%	80.3% \pm 2.4

Face Recognition

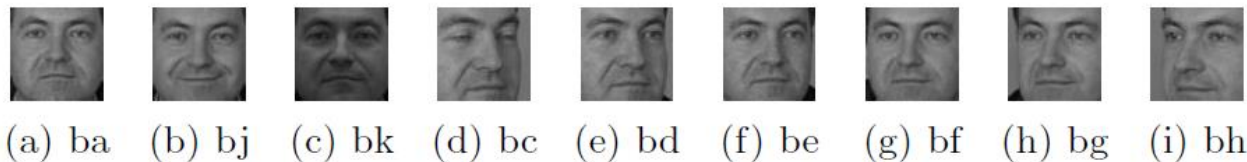


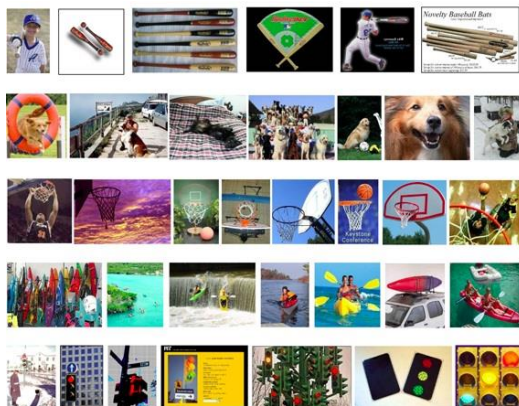
Fig. 4. Samples from the FERET dataset [17]

From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices,
Mehrtash Harandi, Mathieu Salzmann, Richard Hartley, ECCV'14

**Уровень
сложности задач
и контроль
результатов:
Benchmarks**

Benchmarks

Caltech 101



PASCAL VOC



ImageNet



Benchmarks

<http://mscoco.org/>

Microsoft Common Objects in COntext (MS COCO) dataset

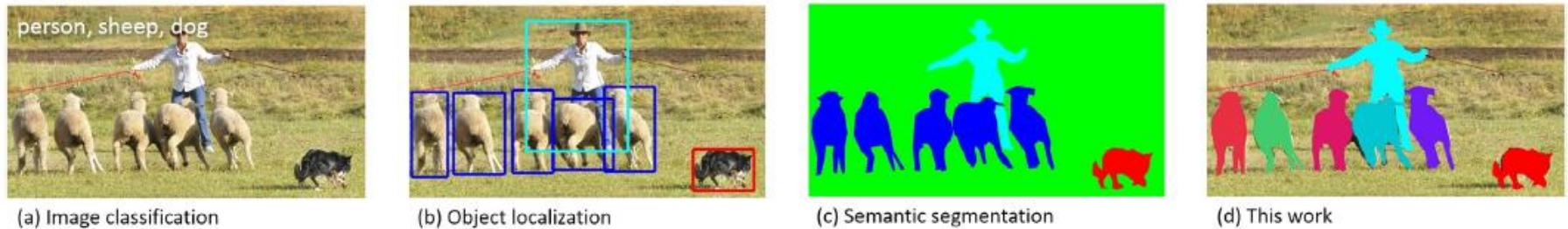


Fig. 1. While previous object recognition datasets have focused on (a) image classification, (b) object bounding box localization or (c) semantic pixel-level segmentation, we focus on (d) segmenting individual object instances. We introduce a large, richly-annotated dataset comprised of images depicting complex everyday scenes of common objects in their natural context

Objects are labeled using per-instance segmentations to aid in precise object localization. Dataset contains photos of 91 objects types that would be easily recognizable by a 4 year old. With a total of 2.5 million labeled instances in 328k images.

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks

<http://mscoco.org/>



Fig. 2. Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images. In this work we focus on challenging non-iconic images.

Objects are labeled using per-instance segmentations to aid in precise object localization. Dataset contains photos of 91 objects types that would be easily recognizable by a 4 year old. With a total of 2.5 million labeled instances in 328k images.

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks

<http://mscoco.org/>

Annotation Pipeline

dog, bottle



(a) Category labeling



(b) Instance spotting



(c) Instance segmentation

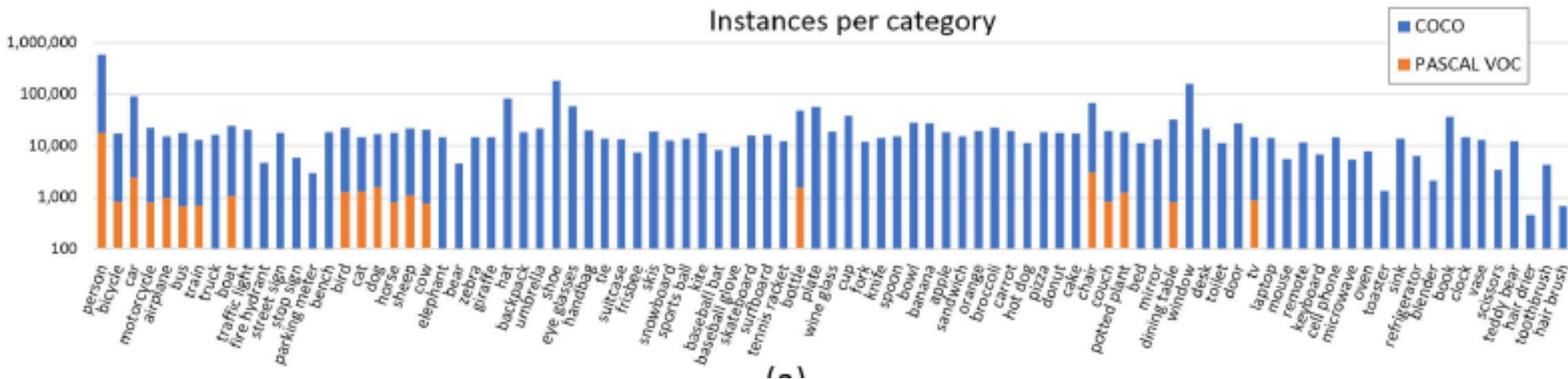
Fig. 3. Our image annotation pipeline is split into 3 primary worker tasks: (a) Labeling the categories present in the image, (b) locating and marking all instances of the labeled categories, and (c) segmenting each object instance.

Microsoft COCO: Common Objects in Context

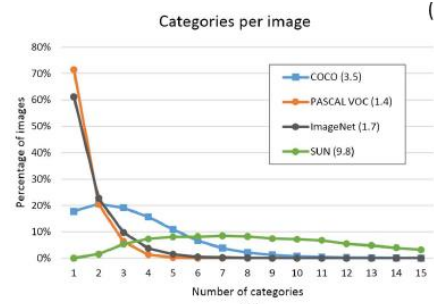
Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks

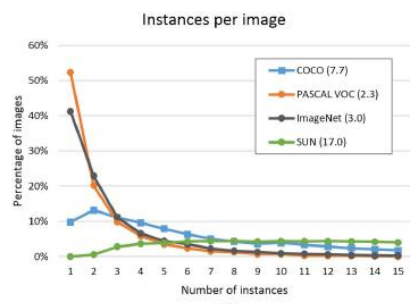
<http://mscoco.org/>



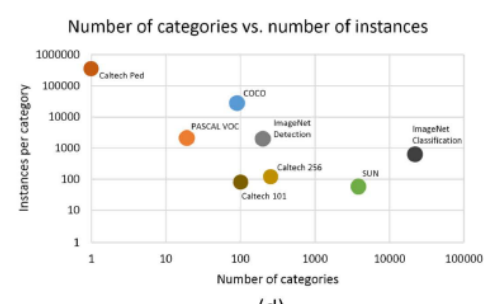
(a)



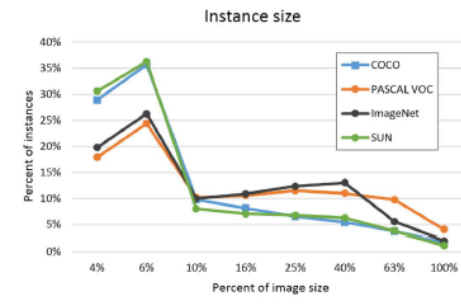
(b)



(c)



(d)



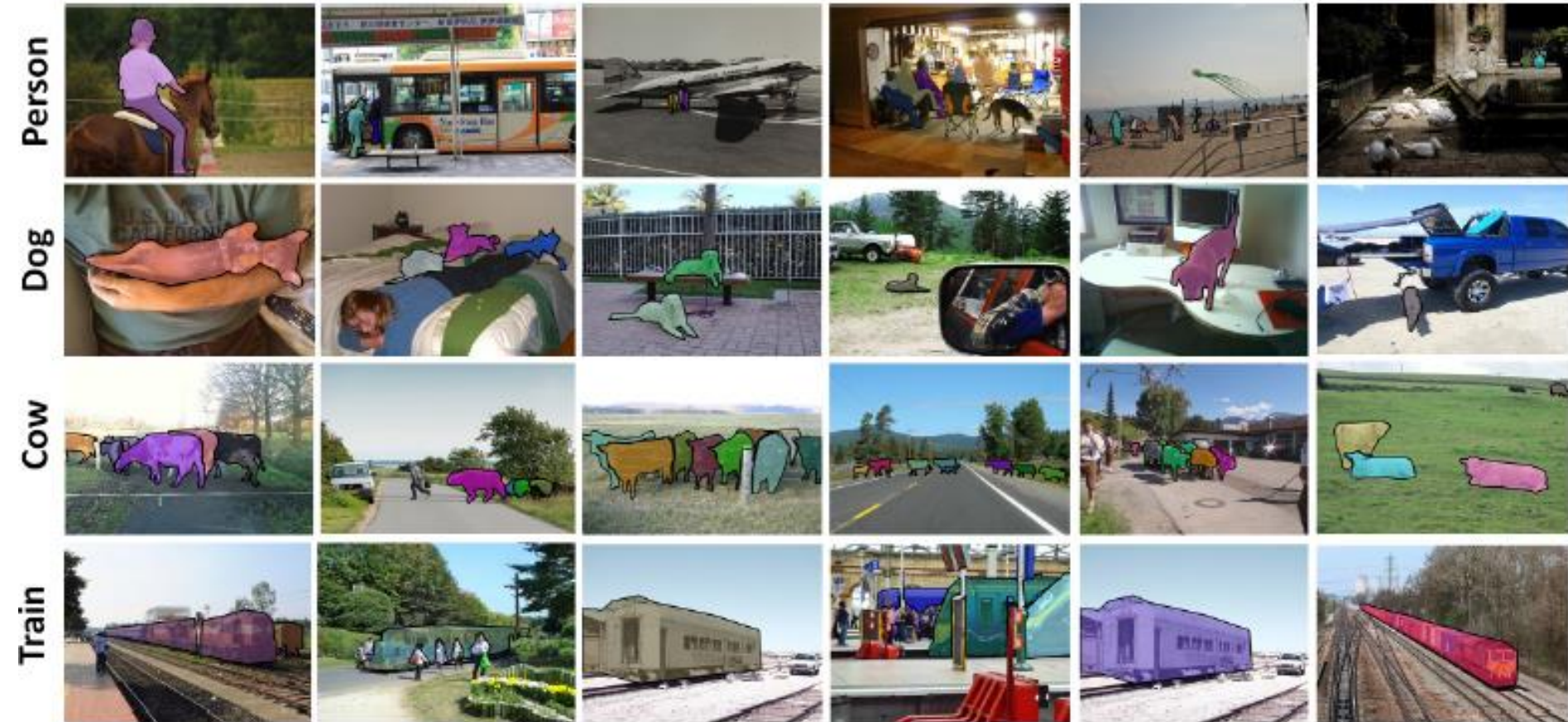
(e)

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks

<http://mscoco.org/>



Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks

<http://mscoco.org/>



Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks, Deformable Parts Model

Bounding-box Detection. We begin by examining the performance of the well studied 20 PASCAL object categories on our dataset. We evaluate two different models.

DPMv5-P: the latest implementation of DPM (release 5) trained on PASCAL VOC 2012.

DPMv5-C: the same implementation trained on COCO (5000 positive and 10000 negative images).

Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: *Object detection with discriminatively trained part-based models*. PAMI 32(9), 1627–1645 (2010)

Girshick, R., Felzenszwalb, P., McAllester, D.: *Discriminatively trained deformable part models, release 5*. PAMI (2012)

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks, Deformable Parts Model

Bounding-box Detection

Table 1. Top: Detection performance evaluated on PASCAL VOC 2012. DPMv5-P is the performance reported by Girshick et al. in VOC release 5. DPMv5-C uses the same implementation, but is trained with MS COCO. Bottom: Performance evaluated on MS COCO for DPM models trained with PASCAL VOC 2012 (DPMv5-P) and MS COCO (DPMv5-C). For DPMv5-C we used 5000 positive and 10000 negative training

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	Avg.
DPMv5-P	45.6	49.0	11.0	11.6	27.2	50.5	43.1	23.6	17.2	23.2	10.7	20.5	42.5	44.5	41.3	8.7	29.0	18.7	40.0	34.5	29.6
DPMv5-C	43.7	50.1	11.8	2.4	21.4	60.1	35.6	16.0	11.4	24.8	5.3	9.4	44.5	41.0	35.8	6.3	28.3	13.3	38.8	36.2	26.8
DPMv5-P	35.1	17.9	3.7	2.3	7	45.4	18.3	8.6	6.3	17	4.8	5.8	35.3	25.4	17.5	4.1	14.5	9.6	31.7	27.9	16.9
DPMv5-C	36.9	20.2	5.7	3.5	6.6	50.3	16.1	12.8	4.5	19.0	9.6	4.0	38.2	29.9	15.9	6.7	13.8	10.4	39.2	37.9	19.1

If we compare the average performance of DPMv5-P on PASCAL VOC and MS COCO, we find that average performance on MS COCO drops by nearly a *factor of 2*, suggesting that MS COCO does include more difficult (non-iconic) images of objects that are partially occluded, amid clutter, etc. We notice a similar drop in performance for the model trained on MS COCO (DPMv5-C).

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks, Deformable Parts Model

Detection Evaluated by Segmentation

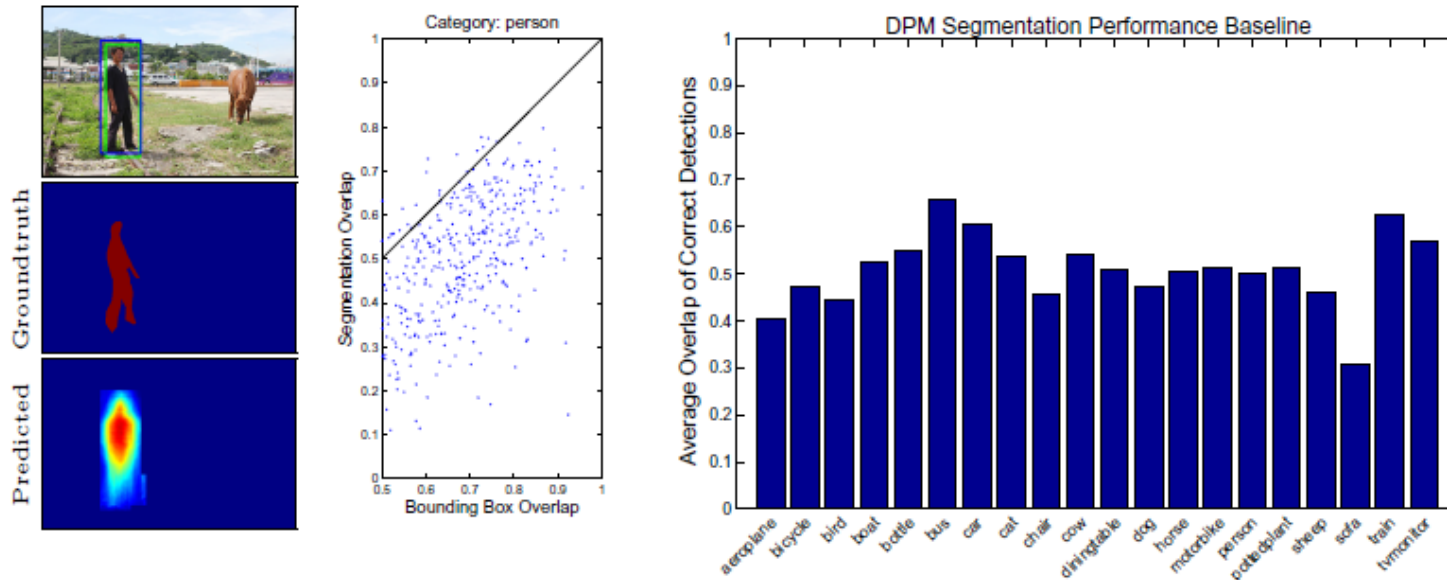


Fig. 9. A predicted segmentation might not recover object detail even though detection and groundtruth bounding boxes overlap well (left). Sampling from the person category illustrates that on a per-instance basis, predicting segmentation from top-down projection of DPM part masks is difficult even for correct detections (center). Averaging over instances for each of the PASCAL VOC categories on our dataset demonstrates that it presents a challenge for object segmentation algorithms (right).

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Benchmarks, Deformable Parts Model

Detection Evaluated by Segmentation

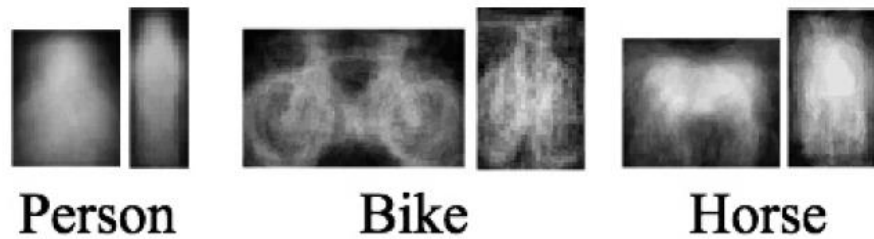


Fig. 7. We visualize our mixture-specific shape masks. We paste thresholded shape masks on each candidate detection to generate candidate segments.

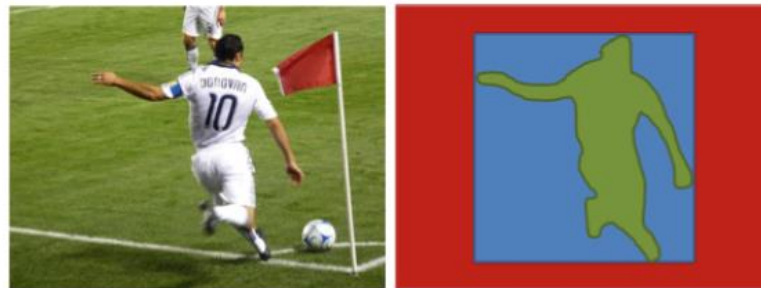


Fig. 8. Evaluating instance detections with segmentation masks versus bounding boxes. Bounding boxes are a particularly crude approximation for articulated objects; in this case, the majority of the pixels in the (blue) tight-fitting bounding-box do not lie on the object. Our (green) instance-level segmentation masks allows for a more accurate measure of object detection and localization.

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin (Cornell), Michael Maire (Caltech), Serge Belongie (Cornell), James Hays (Brown), Pietro **Perona** (Caltech), Deva Ramanan (UC Irvine), Piotr **Dollar** (Microsoft Research), and C. Lawrence **Zitnick** (Microsoft Research), ECCV'14

Заключение I

Задачи компьютерного зрения, доведенные до стадии технологических решений:

- SLAM – технология одновременной реконструкции 3D сцены и оценки положения/параметров движения камеры
- Привязка облаков точек к заданным 3D моделям
- Распознавание характерных элементов городской среды для привязки видеоданных к карте
- Выделение и прослеживание движущихся объектов
- Использование правил анализа динамической сцены для генерации событий и сообщений в системах видеонаблюдения
- Системы автоматического анализа специализированных видеоданных (например, некоторых типов спортивных игр)
- Выделение изменений в сцене
- Обнаружение людей (пешеходов)
- Обнаружение и распознавание лиц
- Поиск по сходству в коллекциях изображений
- Общие успехи в распознавании визуальных образов (Deep Learning)

Заключение II

Задачи компьютерного зрения, находящиеся в стадии фундаментального исследования и далекие от готовых технологических решений:

- Оценка характера поведения групп людей или толпы
- Выделение и прослеживание отдельных людей в толпе
- Ре-идентификация людей при съемке в различных условиях
- Распознавание лиц в сложных условиях съемки, при низком разрешении, при наличии мимики
- Распознавание действий людей, особенно – их взаимодействия различных типов, операций с предметами , взаимодействия с элементами окружающей среды и т.п.
- Реконструкция природных процессов в сценах наблюдения (пожары, наводнения, разрушения, т.п.)
- Семантический поиск по сходству в коллекциях видеоданных
- Автоматическое аннотирование видеоданных с использованием текстовых тегов
- Построение и использование пространственно-временных логик и онтологий для анализа сложных динамических сцен