Large-scale Image Processing

Overview of the sate of the art of publications, software and datasets

E. Ranguelova
Netherlands eScience Center,
Science Park 140 (Matrix 1), 1098 XG Amsterdam, the Netherlands

June 29, 2015

Contents

1 Introduction

The subject of this report is to present an overview of the state of the art in (large scale) computer vision and image processing (CV&IP). Since this is a very large research area, the focus of this report is mostly on three main research questions in CV&IP:

1. **Visual salience:** How can the CV system determine automatically the most visual salient region(s) in an image?
2. **Object/scene identification:** How can the CV system automatically determine weather two images, potentially taken with different cameras under different viewing conditions and transformations, represent the same object/scene?
3. **Object detection and scene understanding:** How can the computer recognise automatically to what visual category the object/scene captured in an image belongs to?

Few of the numerous applications related to the visual saliency (1) are:

- Automatic target detection (see fig. 1)
- Robot navigation using salient objects
- Image and video compression
- Automatic cropping/centering images for display on small portable screens
- Tumor detection in mammograms, etc.

Figure 1: Example of a saliency model detecting the vehicle as being the most salient object ina complex scene.

Few of the numerous applications related to object/scence identification (2) are:

- Stereo and wide-baseline matching
- Image panorama stiching/creation
- Automatic reconstruction of 3D scenes
- Individual wildlife photo-identification (see figs. 2 and 3), etc.

Figure 2: Example species with sufficient spot pattering what could be useful for automated photo-identification: (a) whale shark (with reference area), (b) spotted tree frog, (c) northern quoll, (d) Amazon spotted frog, (e) striped blue crow and (f) mangrove snake.

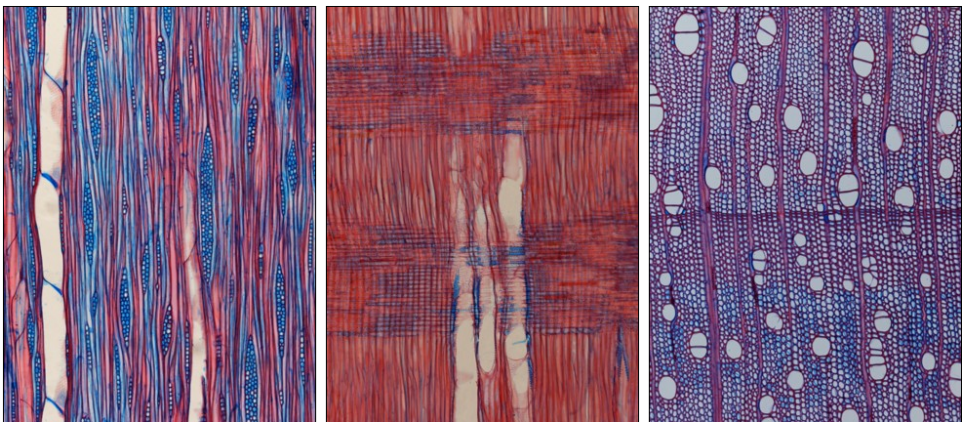

Figure 3: Left to right: tangential, radial and cross (transversal) sections of stained wood Acer.

Automatically understading the semantics of an object/scene (3) have numeral applications like:

- Image search engines

- Organizing photo collections
- Autonomous driving
- Human machine interaction, etc.

This is the most complex and high-level computer vison task, with the goal of making machines see like humans and be able to infer both general principles as well as current situaitons from images. Example of a trained system for scee categorization is shown on fig.4.

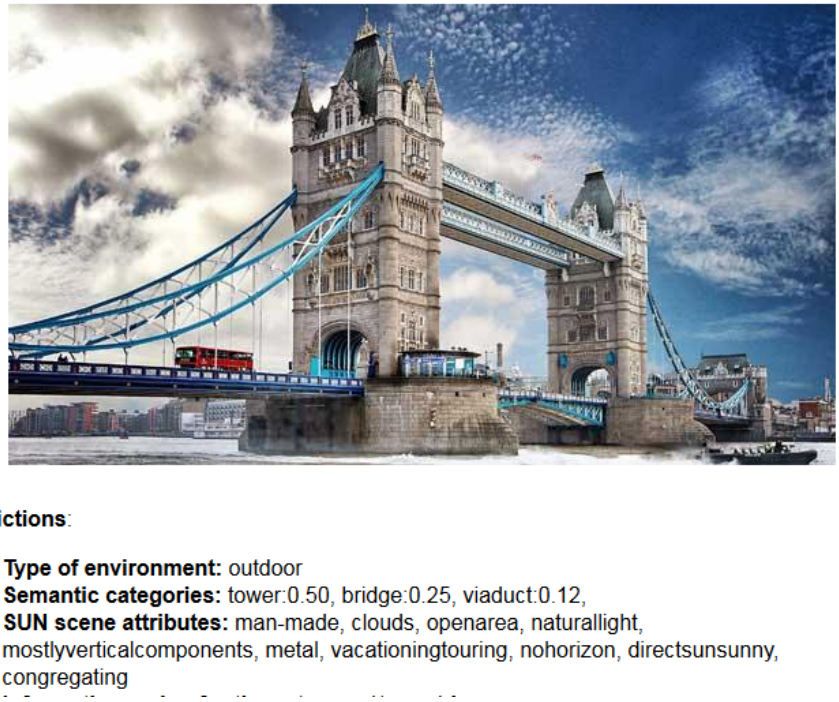

Figure 4: MIT Scene Recognition Demo. `http://places.csail.mit.edu/demo.html`

The overview is by no means complete, it rather tries to summarise the research in the field along the above three questions in the last years. These questions have been chosen having in mind the problems defined in other sciences whcih can be helped by computer science, more precisely by the developments in CV&IP.

The report is structured along the three main products of CV&IP reserach, namely, scientific publications in section 2, software in section 3 and datasets in section 4. Each section gives examples of the work related to each of the three research questions. Some potential scientific aplications are shown in section 5.

The goal of this report is not only to summarize, but is also an attempt to identify suitable "niche" for scientifical domain-driven (applied) research in CV&IP at NLeSc. The conclusions and recommendations are given is section 6.

2 Publications

2.1 Saliency

In [22] the salient object detection is formulated as image segmentation problem. The object is separated from the background on the basis of several features including multi-scale contrast, center-surround histogram and colour spatial distribution for the object description on several levels- locally, regionally and globally. The multi-scale contrast is the local feature, the center-surround histogram is the regional feature and the colour spatial histogram- the global. These features are illustrated on Figure 5.

Figure 5: Examples of salient features. From leftto right: input image, multi-scale contrast, center-surround histogram, colour spatial distribution and binary salient mask by CRF.

A conditional Random Field is trained on these features. For the purposes of this research the autors have compiled a large-scale database, MSRA ([27]), presented in section 4.1.1. The database is publically available, while the software is not. The proposed methods compared to two other algorithms "FG" (fuzzy growing) and "SM" (salient model as computed by the SalientToolbox, described in section 3.1.1). The authors' tends to produce smaller and more focused bounding boxes.

good for automatic cropping? In [1] the authors perform a frequency-domain analysis on five stateof- the-art saliency methods, and compared the spatial frequency content retained from the original image, which is then used in the computation of the saliency maps. This analysis illustrated that the deficiencies of these techniques arise from the use of an inappropriate range of spatial frequencies. Based on this analysis, they presented a frequency-tuned approach of computing saliency in images using low level features of color and luminance. The resulting saliency maps are better suited to salient object segmentation, with higher precision and better recall than the analyzed state-of-the-art techniques.

In [34] the authors address a fundamental problem in saliency detection, namely, the small-scale background structures, which affect the detection. This problem occurs often in natural images. They propose a hierarchical framwork that infers importance values fromimage layers with different scales. The approach is summarized in Figure 6.

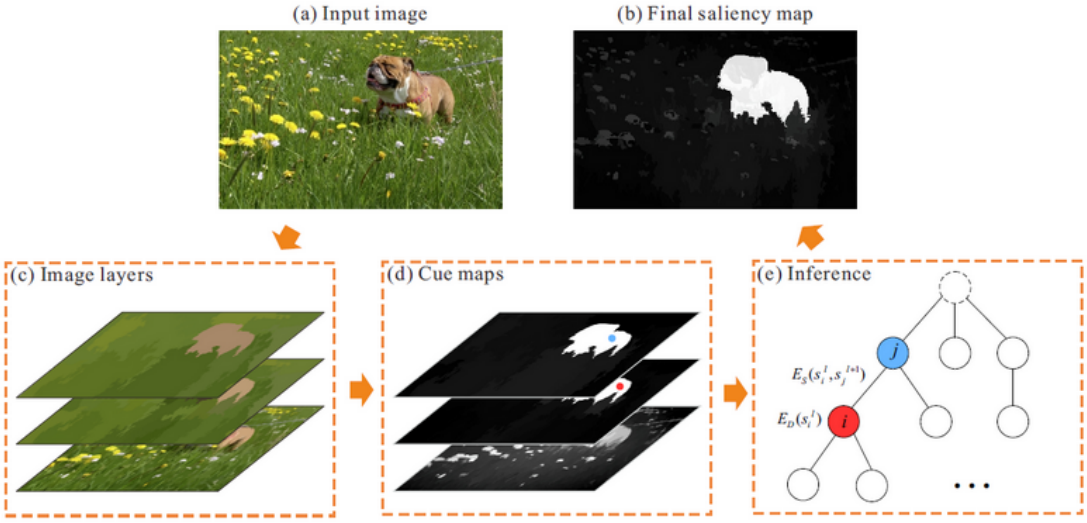

Figure 6: An overview of the hierarchical framework. Three image layers are extracted from the input, and then saliency cues from each of these layers are computed. They are finally fed into a hierarchical model to get the final results.

For the purpose of their research the authors made a new database available to the community, the Complex Scene Saliency dataset (CSSD) and the Extended CSSD (ECSSD), described in Section 4.1.3. The executable of their software is also available from the project link ([33]), but not the source code. The authors report better performence of their method on MSRA-1000 and (E)CSSD datasets comapred to 11 other state-of-the-art methods.

2.2 Salient regions

In [18] colour extension ot the popular *Maximally Stable Extremal Regions (MSER)* detector is proposed. The author calles his detector *Maximally stable colour region (MSCR)*. In comparision on the well-known Visual Geometry Group in Oxford test image sets ([28]) with known homographies to the original MSER detector, the simple colour MSER extension (MSER3) and a colour blob detector, the MSCR performes better in most cases. The executables of the software for MSCR and blob detectors are available at the author's homepage [17].

The MSER detectorhas been also extended in 3D to *Maximally Stable Volumes (MSVs)* in [8]. The MSVs have been used to sucessfully segment 3D medical images and paper fiber networks.

In [16] a structure-guided salient region detector (SGSR) is introduced. It is based on entropy-based saliency theory and shows competitive performance.

In [31] another enhancmenet of MSER is proposal, namely with Canny detector....

2.3 Convolutional Neural Networks

3 Software

The saliency detection, dataset annotation and recognition tools developed by the researchers are often made available to the community.

3.1 Saliency

3.1.1 SaliencyToolbox

The SaliencyToolbox is a collection of Matlab functions and scripts for computing the saliency map for an image, for determining the extent of a proto-object, and for serially scanning the image with the focus of attention.

3.1.2 Frequency-tuned Saliency

The code used in the CVPR 2009 paper "Frequency-tuned Salient Region Detection" ([1]) is accessible through the online presentation of the work at [12].

3.2 Dataset annotation

3.2.1 LabelMe

LabelMe is a WEB-based image annotation tool that allows researchers to label images and share the annotations with the world. The images canbe organizated into collections, which canbe nested. Images can alsobe uploaded into the system andshared.

The LabelMe MATLAB toolbox is used for interaction with the images and annotations in the LabelMe dataset 4.3. The tool is described in this paper [25]. The toolbox also exists in 3D version, LabelMe3D which is described in [5]. There is also a mobile App version and instructions how could the labelling be outsourced using the Amazon Mechanical Turk.

3.2.2 Photo-identification

3.3 Convolutional Neural Networks

With the excellent performance of CNNs on the ImageNet classification dataset and many other recognition tasks, there is a boom of development of software tools implementing deep leanring and CNNs. The tools are often free and open source. At `http://deeplearning.net/software_links/` there is an extensive list of such packages/libraries. Here, only the most popular are presented:

3.3.1 Caffe

Caffe ([14]) is BSD2-Clause license modular framework for deep leanring developed by the Berkeley Vision and Learning Center (BVLC). The framework is a C++ library with Python and MATLAB bindings fr training and deploying general-prurpose CNNs and other deep models on commodity architectures. It is a very popular framework, both in academia and industry due to its speed performance- it can process $60M$ images per day witha single NVIDIA K40 GPU. There is a large community of user and user groups and contributers on GitHub. A technical report for Caffe can be found at it's git repository: `https://github.com/BVLC/Caffe/`.

3.3.2 Torch7

3.3.3 Theano

3.3.4 DeepLearnToolbox

3.3.5 Deeplearning4j

4 Datasets

The annotated segmentation, saliency, object and scenes classification datasets, which are produced and used by the researchers to test the algorithms are often made available to the community.

4.1 Image Saliency Datasets

4.1.1 MSRA

The MSRA Database form the Visual computing group of Microsoft Research Asia [27] is first large-scale labeled dataset made publically available for training and evaluation. It contains two image sets. The first set consists of 20000 images labeled by three users, while the second set consists of 5000 images labeled by nine users. The labeling are available as bounding boxes. Figure 7 illustrates the dataset.

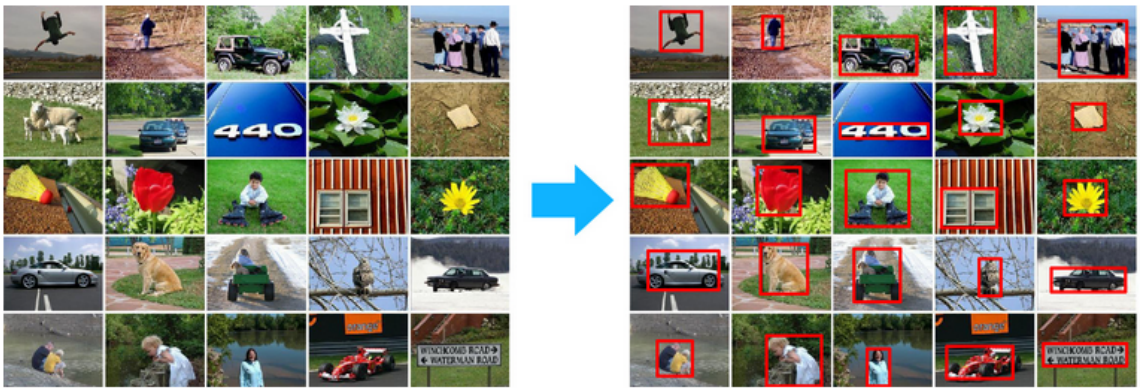

Figure 7: Examples of the MSRA dataset.

The results of the proposed method by the authors of the dataset, have been published in [22].

4.1.2 MSRA10k

This is an extension of the MSRA dataset, which addresses the coarse-grained limitation of the MSRA labeling (bounding boxes). The MSRA10k ([24]) dataset consists of 10000 randomly selected MSRA images for which a pixel-level saliency labeling is available. Figure 8 illustrates the dataset.

Figure 8: Examples of the MSRA 10k dataset. First row: original images with ground truth rectangles from MSRA dataset. Second row: Ground truth with pixel accuracy.

This dataset is used by in a very recent paper in IEEE Transactions on PAMI [6] and [11] (online resources with link to the software).

4.1.3 CSSD and ECSSD

Although images from MSRA-1000 [1] have a large variety in their content, background structures are primarily simple and smooth. To represent the situations that natural images generally fall into, the Complex Scene Saliency Dataset (CSSD) [32] was proposed in [34] with 200 images. They contain diverse patterns in both foreground and background. The labeling has done by five helpers. These images were collected from the BSD300 (later extended to BSD500, [23]), VOC dataset [15] and internet.

Later, the CSSD was extended to a larger dataset (ECSSD) of 1000 images, which includes many semantically meaningful and structurally complex images for evaluation. The images are acquired from the internet and five helpers were asked to produce the ground truth masks. Examples of the images in the dataset can be seen on Figure 9.

4.1.4 DUT-OMRON

The Dalian University of Technology and the Omron Corporation introduced in the DUT-OMRON dataset [35] consisting of 5168, manually selected from more than 140000 images. They are re-sized to $400 \times x$ or $x \times 400$, where $x < 400$. They contain one or more salient objects with relatively complex background. Five people have labeled the pixel-wise ground truth along with bounding box and eye-fixation. The dataset is illustrated on Figure 10. The results of the experiments on the collected dataset were published in [36].

4.1.5 PASCAL-S

Another dataset, which aims at bridging the gap between fixations and salient objects is the PASCAL-S dataset [21] provided by Georgia Tech, Caltech and UCLA. The dataset contains

Figure 9: Examples of the ECSSD dataset.

850 images from the PASCAL 2010 with 12 subjects and 1296 object instances. The images and the code are available for download. The dataset is is illustrated on Figure 11.

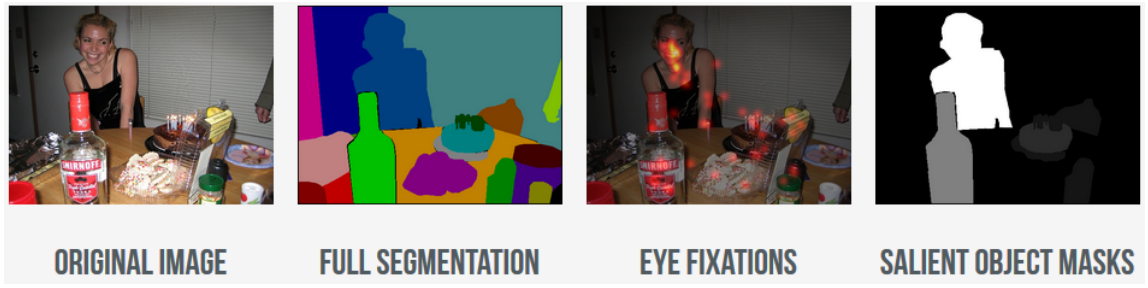

Figure 11: Examples of the PASCAL-S dataset.

The saliency segmentation method and the findings have been published at [20].

4.2 Multimedia Datasets

4.2.1 MSRA-MM

In 2009, the reserachers from Micsosoft Research Asia have released 2 versions of large multimedia datasets- MSRA-MM [29]. MSRA-MM 1.0 consists of two sub-datasets, i.e., an image dataset and a video dataset that are collected from the image and video search engines. For image dataset, there are about 1000 images per query for 68 representative queries based on the log of search engies. There are 65443 images in total. For the video dataset, 165 representative queries have been selected from a log resulting in total of 10277 videos. Due to copyright issues,

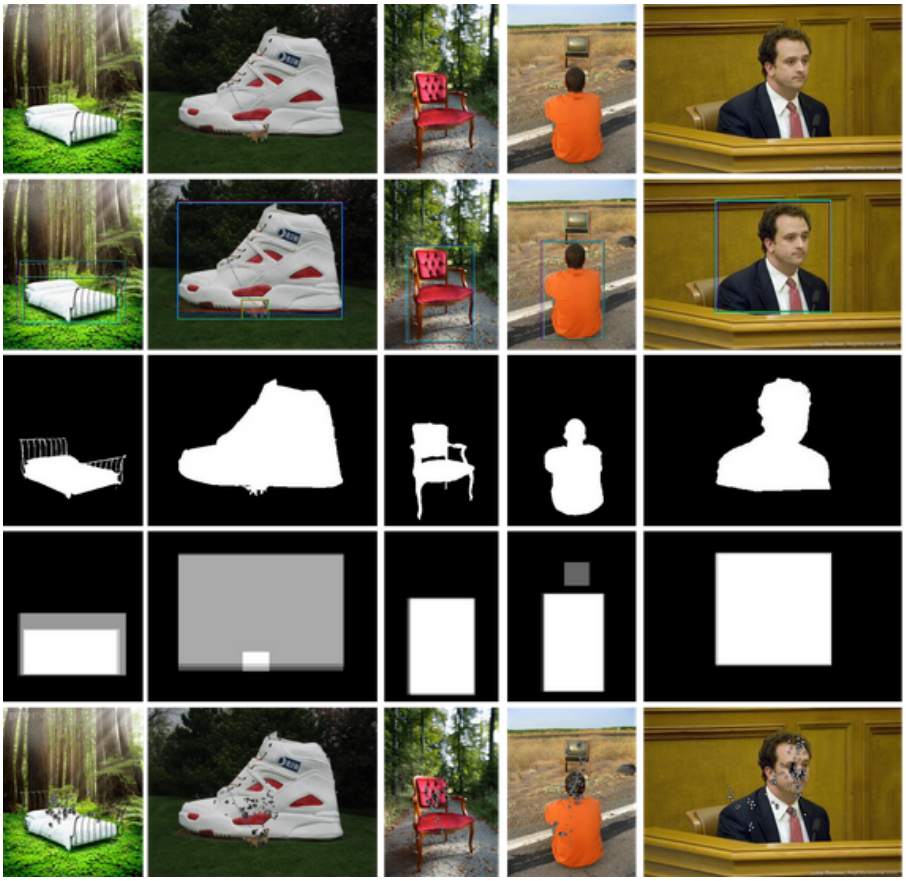

Figure 10: Samples of the DUT-OMRON dataset.From top to bottom: original image, bounding box ground truth, pixel-wise ground truth,average of the five binary masks and eye-fixation ground truth.

the raw image and video data are not available, but only features and annotations are provided. The dataset is explained in detail in a techical report [30].

4.3 Object and Scene Recognition Datasets

4.3.1 MIT-CSAIL

The goal of the MIT-CSAIL dataset [2] is to provide a large set of images of natural scenes (principally office and street scenes), together with manual segmentations/labelings of many types of objects, so that it becomes easier to work on general multi-object detection algorithms. The dataset contains indoor and outdoor objects in officeand urban environments. There are annotations for morethan30objects in cotext in thousands of images and sequences with 2500 annotaited frames. Examples of the dataset are shown in figure 12.

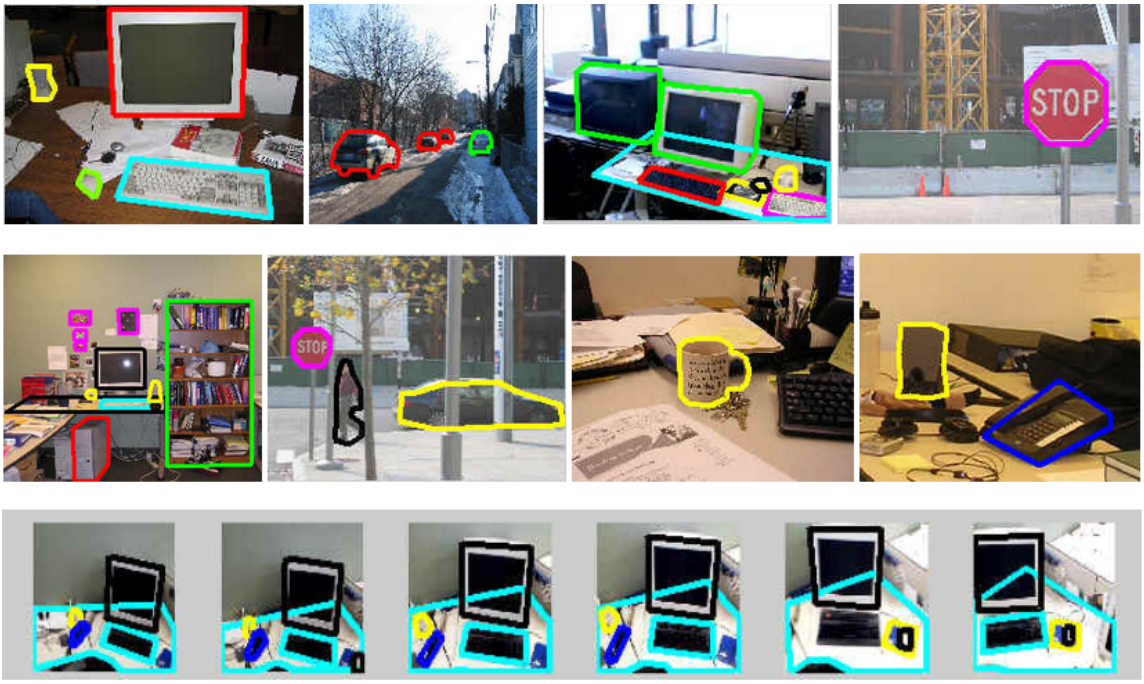

Figure 12: Examples of the MIT-CSAIL dataset.

4.3.2 LabelMe

An online annotated dataset incrementally filled up by users can be downloaded using the LabelME tool3.2.1 from [4]. The LabelMe3D dataset contains labelled images of many everyday scenes and object categories in absolute real world 3D coordinates. The toolboxes designed to work with these datasets are described in 3.2.1.

4.3.3 SUN

Another large dataset of annotated images covering large vaiety of scenes, places and objects within, provided also by the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT, is the SUN dataset, [13]. The SUN2012 contains 16873 images and SUN contains currently 131067 images, 908 Scene categories and 4479 object categories with more than $310k$ segmented objects. The SUN397 benchmark for scene classification can de used incuding code, precomputed features etc. The SUN dataset can bee dowloaded also with the LabelMe MATLAB Toolbox (3.2.1). The publication about the SUN dataset is [3].

4.3.4 Places

One of the largest annotated datasets for scene recognition is the Places dataset (also by CSAIL, MIT), [7]. It contains almost 2.5 million images in 205 scene categories. Along with the dataset, one can access the Places-CNNs, the convolutional neural networks trained on Places, DrawCNN- a visualization of the units' connections for the CNN, the online recognition demo and some sample MATLAB code forusing the synthetic receptive field of unit to segment image and visualize the activated regions. The pulications where the dataset is described are [9, 10].

5 Applications

There are numerous domains where large scale image processing (LSIP) techniques can be applied. For the scope of this report we are interested in scientific domains where the LSIP can automate tasks performed by the researchers.

5.1 Animal biometrics

In [19], Kuehl and Burghardt give overview of the methodologies and trends in the emerging field of *animal biometrics*. It is an exciting field operating at the intersection between pattern recogniiton, ecology andinformation sciences. The subject of the field is to produce computerized systems for phenotypic measurement and interpretation. The main questions for which such systems helps to find the answers to are: how to profile species, individuals and animal behaviour by representing phenotypic appearance. Figure 13 illustrates the main components of a biometric system. That system parts can either be connected directly on-site or remotely via networks. Each of the components is illustrated, using individual African penguin recognition by spot pattern as an example. Acquisition: automatic or semi-automatic collection of images or video from fixed field cameras, observers or the general public. Detection: the use of computer algorithms to search the images to find those that contain the biometric entity of interest and then to extract relevant information about that entity (e.g., the chest spots of a penguin). Storage: the extracted data on the entity is reduced to a compact mathematical form that can be stored in a suitable database. Matching: the mathematical data on the entity are then compared with other data already stored in the database to find matches that enable the individual or the behavior to be identified, using methods akin to the matching of fingerprints to identify humans. Interfacing: presenting the output of the biometric system to a user or software system for further analysis.

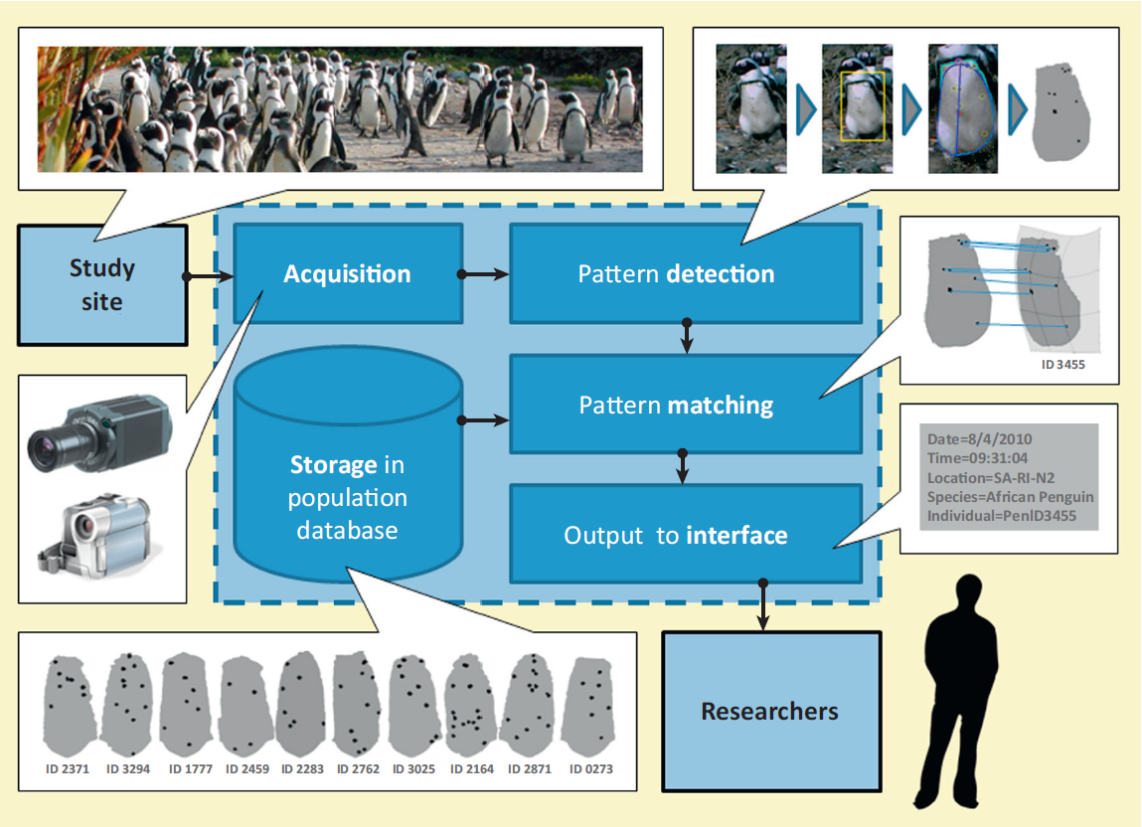

Figure 13: Main components of an animal biometric system. This flowchart summarizes how information from a study site is measured and interpreted for the researcher by an animal biometric system.

Animal biometrics is important field not only for ecological researchers, but for thegeneral public. For example, in [26], a biometric system for face recognition of pet animals (mainly dogs) have been developed.

6 Conclusions

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Ssstrunk. Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1597 – 1604, 2009.

[2] Kevin Murphy Antonio Torralba and William Freeman. The mit-csail database of objects and scenes. `http://web.mit.edu/torralba/www/database.html`. [Online; accessed 18 June 2015].

[3] J. Xiao at al. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010.

[4] A. Torralba et al. B. Russell. Labelme. http://labelme2.csail.mit.edu/Release3.0/browserTools/php/dataset.php.

[5] B.C.Russell and A. Torralba. Building a Database of 3D Scenes from User Annotations. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 2009.

[6] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.

[7] MIT CSAIL. Places, the scence recognition database. http://places.csail.mit.edu/.

[8] Michael Donoser and Horst Bischof. 3d segmentation by maximally stable volumes (msvs). In *ICPR (1)*, pages 63–66. IEEE Computer Society, 2006.

[9] B. Zhou et al. Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

[10] B. Zhou et al. Object detectors emerge in deep scene cnns. *International Conference on Learning Representations (ICLR)*, 2015.

[11] Ming-Ming Cheng et al. Global contrast based salient region detection. http://mmcheng.net/zh/salobj. [Online; accessed 13 April 2015].

[12] Radhakrishna Achanta et al. Frequency-tuned salient region detection. http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/index.html. [Online; accessed 13 April 2015].

[13] Xiao et al. Sun database. http://groups.csail.mit.edu/vision/SUN/.

[14] Yangqing Jia et al. Caffe. http://caffe.berkeleyvision.org/. [Online; accessed 29 June 2015].

[15] Mark Everingham. The pascal visual object classes homepage. http://pascallin.ecs.soton.ac.uk/challenges/VOC/. [Online; accessed 14 April 2015].

[16] Shufei Fan and Frank Ferrie. Structure guided salient region detector. In *In Proceedings of British Machine Vision Conference*, pages 423–432, 2008.

[17] Per-Erik Forssen. Maximally stable colour regions. http://www.cs.ubc.ca/~perfo/mscr/. [Online; accessed 21 April 2015].

[18] Per-Erik Forssén. Maximally stable colour regions for recognition and matching. In *CVPR*, 2007.

[19] H. Kuehl and T. Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in Ecology & Evolution*, 28:432–441, 2013.

[20] Jian Li, M. D. Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):996–1010, 2013.

[21] Yin Li. The pascal-s dataset. http://cbi.gatech.edu/salobj/. [Online; accessed 14 April 2015].

[22] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1–8, 2007.

[23] David Martin. Bsd300/500: The berkeley segmentation dataset and benchmark. `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`. [Online; accessed 14 April 2015].

[24] M.M.Cheng. Msra10k: Pixel accurate salient object labeling for 10 000 images from msra dataset. `http://mmcheng.net/msra10k/`. [Online; accessed 13 April 2015].

[25] B.C. Russell, A. Torralba, K.P. Murhy, and W.T.Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008.

[26] Santosh Kumarand Sanjay Singh. Biometric recognition for pet animal. *Journal of Software Engineering and Applications*, 7:470–482, 2014.

[27] Microsoft Research Visual Computing Group. Msra salient object database. `http://research.microsoft.com/en-us/um/people/jiansun/salientobject/salient_object.htm`. [Online; accessed 13 April 2015].

[28] Oxford Visual Geometry Group. Affine covariant regions. `http://www.robots.ox.ac.uk/~vgg/research/affine/`. [Online; accessed 21 April 2015].

[29] Meng Wang. Msra-mm - msr asia internet multimedia dataset 1.0 and 2.0. `http://research.microsoft.com/en-us/projects/msrammdata/`. [Online; accessed 14 April 2015].

[30] Meng Wang, Linjun Yang, and Xian-Sheng Hua. Msra-mm: Bridging research and industrial societies for multimedia information retrieval. Technical Report MSR-TR-2009-30, Microsoft, March 2009.

[31] Sh. Wang, W. Wang, D Liu, F. Gu, and B.B.Dickson. Enhanced maximally stable extremal regions with canny detector and applicationin image classification. *Journal of Computational Information Systems*, 10(14):6093–6100, 2014.

[32] Qiong Yan. Cssd: Complex scene saliency dataset. `http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html`. [Online; accessed 14 April 2015].

[33] Qiong Yan. Ecssd: Extended complex scene saliency dataset. `http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html`. [Online; accessed 13 April 2015].

[34] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical Saliency Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, pages 1155 – 1162, 2013.

[35] Chuan Yang. The dut-omron image dataset. `http://202.118.75.4/lu/DUT-OMRON/index.htm`. [Online; accessed 14 April 2015].

[36] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3166–3173. IEEE, 2013.