Unsupervised semantic discovery through visual patterns detection

Francesco Pelosin, Andrea Gasparetto, Andrea Albarelli, and Andrea Torsello

Ca' Foscari University, Venice, Italy
{francesco.pelosin, andrea.gasparetto, andrea.albarelli, andrea.torsello}@unive.it

Abstract. We propose a new fast fully unsupervised method to discover semantic patterns. Our algorithm is able to hierarchically find visual categories and produce a segmentation mask where previous methods fail. Through the modeling of what is a visual pattern in an image, we introduce the notion of "semantic levels" and devise a conceptual framework along with measures and a dedicated benchmark dataset for future comparisons. Our algorithm is composed by two phases. A filtering phase, which selects semantic hotspots captured as clusters of *splashes*, star graphs in descriptor space encoding feature repetitions; then a clustering phase which propagates the semantic properties of the hotspots on a superpixels basis. We provide both qualitative and quantitative experimental validation, achieving optimal results in terms of robustness to noise and semantic consistency. We also made code and dataset publicly available

Keywords: visual-pattern-detection · semantic-discovery · cosegmentation

1 Introduction

The extraction of semantic categories from images is a fundamental task in image understanding [18,5,29]. While the task is one that has been widely investigated in the community, most approaches are supervised, making use of labels to detect semantic categories [2]. Comparatively less effort has been put to investigate automatic procedures which enable an intelligent system to learn autonomously extrapolating visual semantic categories without any *a priori* knowledge of the context.

We observe the fact that in order to define what a visual pattern is, we need to define a scale of analysis (objects, parts of objects etc.). We call these scales *semantic levels* of the real world. Unfortunately most influential models arising from deep learning approaches still show a limited ability over scale invariance [25,13] which instead is common in nature. In fact, we don't really care much about scale, orientation or partial observability in the semantic world. For us, it is way more important to preserve an "internal representation" that matches reality [6,17].

Fig. 1. A real world example of unsupervised segmentation of a grocery shelf. Our method can automatically discover both low-level coherent patterns (brands, flavor images and logos) and high-level compound objects (multi-packs and bricks) by controlling the semantic level of the detection and segmentation process.

Our method leverages repetitions to capture the internal representation in the real world and then extrapolates categories at a specific semantic level. We do this without continuous geometrical constraints on the visual pattern disposition, which is common among other methodologies [21,21,10,22,8].

We also do not constrain ourselves to find only one visual pattern, which is another very common assumption. Indeed, what if the image has more than one visual pattern? One can observe that this is *always* the case. Each visual repetition can be hierarchically decomposed in its smaller parts which, in turn, repeat over different semantic levels. This peculiar observation allow our work to contribute to the community as follows:

- A new pipeline able to capture semantic categories with the ability to hierarchically span over semantic levels.
- A better conceptual framework to evaluate analogous works through the introduction of the semantic levels notion along with a new metric.
- A new benchmark dataset of 208 labelled images for visual repetition detection.

Code, dataset and notebooks are public and live: https://git.io/JT6UZ.

2 Related Works

Several works have been proposed to tackle visual pattern discovery and detection. While the paper by Leung and Malik [11] could be considered seminal, many other works build on their basic approach, working by detecting contiguous structures of similar patches by knowing the window size enclosing the distinctive pattern.

One common procedure in order to describe what a pattern is, consists to first extract descriptive features such as SIFT to perform a clustering in the feature space and then model the group disposition over the image by exploiting geometrical constraints, as in [21] and [4], or by relying only on appearance, as in [7,14,27].

The geometrical modeling of the repetitions usually is done by fitting a planar 2-D lattice, or a deformation of it [20], through RANSAC procedures as in [23]

[21] or even by exploiting the mathematical theory of crystallographic groups as in [15]. Shechtman and Irani [24], also exploited an active learning environment to detect visual patterns in a semi-supervised fashion. For example Cheng et al. [3] use input scribbles performed by a human to guide detection and extraction of such repeated elements, while Huberman and Fattal [9] ask the user to detect an object instance and then the detection is performed by exploiting correlation of patches near the input area.

Recently, as a result of the new wave of AI-driven Computer Vision, a number of Deep Leaning based approaches emerged, in particular Lettry et al. [10] argued that filter activation in a model such as AlexNet can be exploited in order to find regions of repeated elements over the image, thanks to the fact that filters over different layers show regularity in the activations when convolved with the repeated elements of the image. On top of the latter work, Rodríguez-Pardo et al. [22] proposed a modification to perform the texture synthesis step.

A brief survey of visual pattern discovery in both video and image data, up to 2013, is given by Wang et al. [28], unfortunately after that it seems that the computer vision community lost interest over this challenging problem. We point out that all the aforementioned methods look for *only one* particular visual repetition except for [14] that can be considered the most direct competitor and the main benchmark against which to compare our results.

3 Method Description

3.1 Features Localization and Extraction

Visual repetitive patterns can be detected through the boundaries of such visual elements. The first step of our algorithm consists in the extraction of a set $\mathcal{C}$ of *keypoints* indicating a position $\mathbf{c}_j$ in the image. To extract keypoints, we opted for the Canny algorithm, for its simplicity and efficiency, although more recent and better edge extractor could be used [16] to have a better overall procedure.

A descriptor d_j is then computed for each selected $\mathbf{c}_j \in \mathcal{C}$ thus obtaining a *descriptor set* $\mathcal{D}$. In particular, we adopted the DAISY algorithm because of its appealing dense matching properties that nicely fit our scenario. Again, here we can replace this module of the pipeline with something more advanced such as [19] at the cost of some computational time.

3.2 Semantic Hot Spots Detection

In order to detect self-similar patterns in the image we start by associating the k most similar descriptors for each descriptor $\mathbf{d}_j$. We can visualize this data structure as a star subgraph with k endpoints called *splash* "centered" on descriptor $\mathbf{d}_j$. Figure 2 (a) shows one.

Splashes potentially encode repeated patterns in the image and similar patterns are then represented by similar splashes. The next step consists in separating these splashes from those that encode noise only, this is accomplished through an accumulator space.

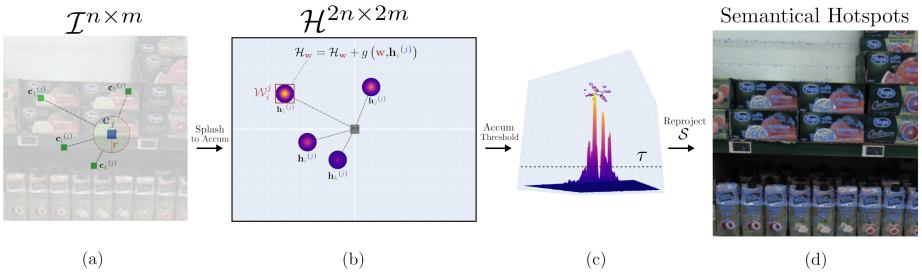

Fig. 2. (a) A splash in the image space with center in the keypoint $\mathbf{c}_j$. (b) $\mathcal{H}$, with the superimposed splash at the center, you can note the different levels of the vote ordered by endpoint importance i.e. descriptor similarity. (c) 3D projection showing the gaussian-like formations and the thresholding procedure of $\mathcal{H}$. (d) Backprojection through the set $\mathcal{S}$.

In particular, we consider a 2-D *accumulator space* $\mathcal{H}$ of size double the image. We then superimpose each splash on the space $\mathcal{H}$ and cast k votes as shown in Figure 2 (b). In order to take into account the noise present in the splashes, we adopt a gaussian vote-casting procedure $g(\cdot)$. Similar superimposed splashes contribute to similar locations on the accumulator space, resulting in peak formations (Figure 2 (c)). We summarize the voting procedure as follows:

$$\mathcal{H}_{\mathbf{w}} = \mathcal{H}_{\mathbf{w}} + g(\mathbf{w}, \mathbf{h}_i^{(j)}) \tag{1}$$

where $\mathbf{h}_i^{(j)}$ is the k-th splash endpoint of descriptor $\mathbf{d}_j$ in accumulator coordinates and $\mathbf{w}$ is the size of the gaussian vote. We filter all the regions in $\mathcal{H}$ which are above a certain *threshold* τ, to get a set $\mathcal{S}$ of the locations corresponding to the peaks in $\mathcal{H}$. The τ parameter acts as a coarse filter and is not a critical parameter to the overall pipeline. A sufficient value is to set it to $0.05 \cdot max(\mathcal{H})$. Lastly, in order to visualize the semantic hotspots in the image plane we map splash locations between $\mathcal{H}$ and the image plane by means of a *backtracking structure* $\mathcal{V}$.

In summary, the key insight here is that similar visual regions share similar splashes, we discern noisy splashes from representative splashes through an auxiliary structure, namely an accumulator. We then identify and backtrack in the image plane the semantic hotspots that are candidate points part of a visual repetition.

3.3 Semantic Categories Definition and Extraction

While the first part previously described, acts as a filter for noisy keypoints allowing to obtain a good pool of candidates, we now transform the problem of finding visual categories in a problem of dense subgraphs extraction.

We enclose semantic hotspots in superpixels, this extends the semantic significance of such identified points to a broader, but coherent, area. To do so we

Algorithm 1 Semantic categories extraction algorithm

Require: G weighted undirected graph
  $i = 0$
  $s^* = -\inf$
  $K^* = \emptyset$
  while G_i is not fully disconnected **do**
    $i = i + 1$
    Compute G_i by corroding each edge with the minimum edge weight
    Extract the set K_i of all connected components in G_i
    $s(G_i, K_i) = \sum_{k \in K_i} \mu(k) - \alpha |K_i|$
    if $s(G_i, K_i) > s^*$ **then**
      $s^* = s(G_i, K_i)$
      $K^* = K_i$
  return s^*, K^*

use the SLIC [1] algorithm which is a simple and one of the fastest approaches to extract superpixels as pointed out in this recent survey [26]. Here comes the choice of the cardinality of the *superpixels* $\mathcal{P}$ to extract. This is the second and most fundamental parameter that will allow us to span over different semantic levels.

Once the superpixels have been extracted, let $\mathcal{G}$ be an *undirected weighted graph* where each node correspond to a superpixel $p \in \mathcal{P}$. In order to put edges between graph nodes (i.e. two superpixels), we exploit the splashes origin and endpoints. In particular the strenght of the connection between two vertices in $\mathcal{G}$ is calculated with the number of splashes endpoints falling between the two in a mutual coherent way. So to put a weight of 1 between two nodes we need exactly 2 splashes endpoints falling with both origin and end point in the two candidate superpixels.

With this construction scheme, the graph has clear dense subraphs formations. Therefore, the last part simply computes a partition of $\mathcal{G}$ where each connected component correspond to a cluster of similar superpixels. In oder to achieve such objective we optimize a function that is maximized when we partition the graph to represent so. To this end we define the following *density score* that given G and a set K of connected components captures the optimality of the clustering:

$$s(G, K) = \sum_{k \in K} \mu(k) - \alpha |K| \tag{2}$$

where $\mu(k)$ is a function that computes the average edge weight in a undirected weighted graph.

The first term, in the score function, assign a high vote if each connected component is dense. While the second term acts as a regulator for the number of connected components. We also added a weighting factor α to better adjust the procedure. As a proxy to maximize this function we devised an *iterative algorithm* reported in Algorithm 1 based on graph corrosion and with temporal complexity

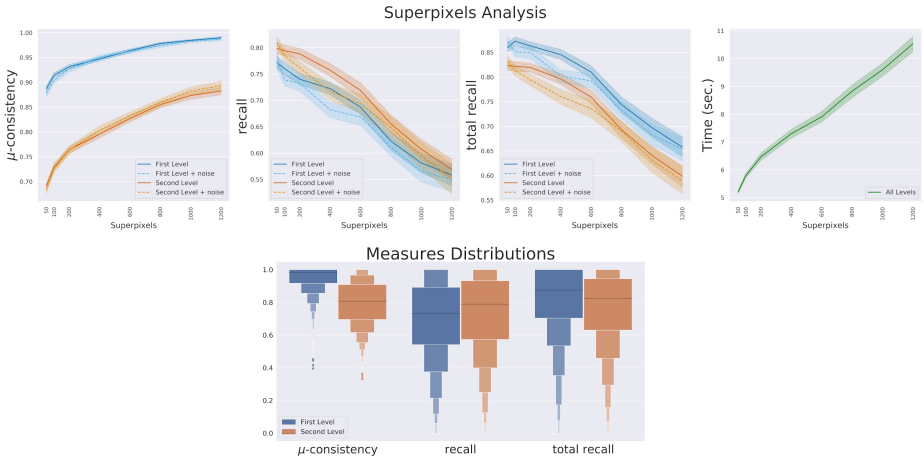

Fig. 3. (top) Analysis of measures as the number of superpixels $|\mathcal{P}|$ retrieved varies. The rightmost figure shows the running time of the algorithm. We repeated the experiments with the noisy version of the dataset but report only the mean since variation is almost equal to the original one. (bottom) Distributions of the measures for the two semantic levels, by varying the two main parameters r and $|\mathcal{P}|$.

of $O(|E|^2 + |E|\,|V|)$. At each step the procedure corrupts the graph edges by the minimum edge weight of G. For each corroded version of the graph that we call *partition*, we compute s to capture the density. Finally the algorithm selects the corroded graph partition which maximizes the s and subsequently extracts the nodes groups.

In brevity we first enclose semantic hotspots in superpixels and consider each one as a node of a weighted graph. We then put edges with weight proportional to the number of splashes falling between two superpixels. This results in a graph with clear dense subgraphs formations that correspond to superpixels clusters i.e. *semantic categories*. The semantic categories detection translates in the extraction of dense subgraphs. To this end we devised an iterative algorithm based on graph corrosion where we let the procedure select the corroded graph partition that filters noisy edges and let dense subgraphs emerge. We do so by maximizing score that captures the density of each connected component.

4 Experiments

Dataset As we introduced in Section 1 one of the aims of this work is to provide a better comparative framework for visual pattern detection. To do so we created a public dataset by taking 104 pictures of store shelves. Each picture has been took with a 5mpx camera with approximatively the same visual conditions. We also rectified the images to eliminate visual distortions.

We manually segmented and labeled each repeating product in two different semantic levels. In the **first semantic level** *products made by the same company*

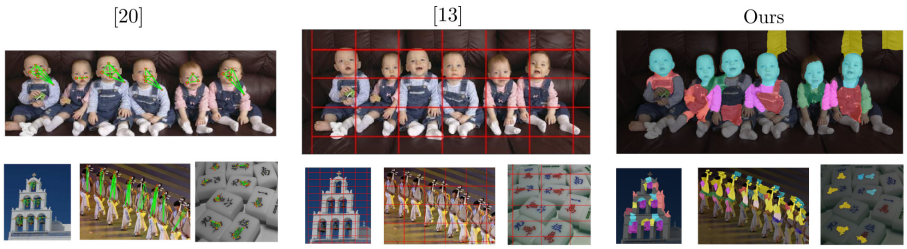

Fig. 4. Qualitative comparison between [14], [10] and our algorithm. Our method detects and segments more than one pattern and does not constrain itself to a particular geometrical disposition.

share the same label. In the **second semantic level** visual repetitions consist in the *exact identical products instances*. In total the dataset is composed by 208 ground truth images, half in the first level and the rest for the second one.

μ-consistency We devised a new measure that captures the semantic consistency of a detected pattern that is a proxy of the average precision of detection.

In fact, we want to be sure that all pattern instances fall on similar ground truth objects. First we introduce the concept of semantic consistency for a particular pattern $\mathbf{p}$. Let $\mathbf{P}$ be the set of patterns discovered by the algorithm. Each pattern $\mathbf{p}$ contains several instances $\mathbf{p}_i$. $\mathbf{L}$ is the set of ground truth categories, each ground truth category $\mathbf{l}$ contain several objects instances $\mathbf{l}_i$. Let us define $\mathbf{t}_p$ as the vector of ground truth labels touched by all instances of $\mathbf{p}$. We say that $\mathbf{p}$ is consistent if all its instances $\mathbf{p}_i, i = 0 \dots |\mathbf{p}|$ fall on ground truth regions sharing the same label. In this case $\mathbf{t}_p$ would be uniform and we consider $\mathbf{p}$ a good detection. The worst scenario is when given a pattern $\mathbf{p}$ every $\mathbf{p}_i$ falls on objects with different label $\mathbf{l}$ i.e. all the values in $\mathbf{t}_p$ are different.

To get an estimate of the overall consistency of the proposed detection, we average the consistency for each $\mathbf{p} \in \mathbf{P}$ giving us:

$$\mu\text{-consistency} = \frac{1}{|\mathbf{P}|} \sum_{\mathbf{p} \in \mathbf{P}} \frac{|\text{mode}\,(\mathbf{t}_p)|}{|\mathbf{t}_p|} \tag{3}$$

recall The second measure is the classical recall over the objects retrieved by the algorithm. Since our object detector outputs more than one pattern we average the recall for each ground truth label by taking the best fitting pattern.

$$\frac{1}{|\mathbf{L}|} \sum_{\mathbf{l} \in \mathbf{L}} \max_{\mathbf{p} \in \mathbf{P}} \ \text{recall}\ (\mathbf{p}, \mathbf{l}) \tag{4}$$

The last measure is the **total recall**, here we consider a hit if any of the pattern falls in a labeled region. In general we expect this to be higher than the recall.

We report the summary performances in Figure 4. As can be seen the algorithm achieves a very high μ-consistency while still able to retrieve the majority of the ground truth patterns in both levels.

One can observe in Figure 3 an inverse behaviour between recall and consistency as the number of superpixels retrieved grows. This is expected since less superpixels means bigger patterns, therefore it is more likely to retrieve more ground truth patterns.

In order to study the robustness we repeated the same experiments with an altered version of our dataset. In particular for each image we applied one of the folling corruptions: Additive Gaussian Noise ($scale = 0.1 * 255$), Gaussian Blur ($\sigma = 3$), Spline Distortions (grid affine), Brightness (+100), and Linear Contrast (1.5).

Qualitative Validation Firstly we begin the comparison by commenting on [14]. One can observe that our approach has a significant advantage in terms of how the visual pattern is modeled. While the authors model visual repetitions as geometrical artifacts associating points, we output a higher order representation of the visual pattern. Indeed the capability to provide a segmentation mask of the repeated instance region together the ability to span over different levels unlocks a wider range of use cases and applications.

As qualitative comparison we also added the latest (and only) deep learning based methodology [10] we found. This methodology is only able to find a single instance of visual pattern, namely the most frequent and most significant with respect to the filters weights. This means that the detection strongly depends from the training set of the CNN backbone, while our algorithm is fully unsupervised and data agnostic.

Quantitative Validation We compared quantitatively our method against [14] that constitutes, to the best of our knowledge, the only work developed able to detect more than one visual pattern. We recreated the experimental settings of the authors by using the Face dataset [12] as benchmark achieving 1.00 precision vs. 0.98 of [14] and 0.77 in recall vs. and 0.63. We considered a miss on the object retrieval task, if more than 20% of a pattern total area falls outside from the ground truth. The parameter used were $|\mathcal{C}| = 9000$, $k = 15$, $r = 30$, $\tau = 5$, $|\mathcal{P}| = 150$

5 Conclusions

With this paper we introduced a fast and unsupervised method addressing the problem of finding semantic categories by detecting consistent visual pattern repetitions at a given scale. The proposed pipeline hierarchically detects self-similar regions represented by a segmentation mask.

As we demonstrated in the experimental evaluation, our approach retrieves more than one pattern and achieves better performances with respect to competitors methods. We also introduce the concept of *semantic levels* endowed with

a dedicated dataset and a new metric to provide to other researchers tools to evaluate the consistency of their approaches.

5.1 Acknowledgments

We would like to express our gratitude to Alessandro Torcinovich and Filippo Bergamasco for their suggestions to improve the work. We also thank Mattia Mantoan for his work to produce the dataset labeling.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. (2012) 5
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV) (2018) 1
3. Cheng, M., Zhang, F., Mitra, N.J., Huang, X., Hu, S.: Repfinder: finding approximately repeated scene elements for image editing. ACM Trans. Graph. (2010) 3
4. Chum, O., Matas, J.: Unsupervised discovery of co-occurrence in sparse high dimensional data. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010 (2010) 2
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Conference on computer vision and pattern recognition CVPR (2016) 1
6. DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition? Neuron (2012) 1
7. Doubek, P., Matas, J., Perdoch, M., Chum, O.: Image matching and retrieval by repetitive patterns. In: 20th International Conference on Pattern Recognition, ICPR 2010 (2010) 2
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016) 2
9. Huberman, I., Fattal, R.: Detecting repeating objects using patch correlation analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016) 3
10. Lettry, L., Perdoch, M., Vanhoey, K., Gool, L.V.: Repeated pattern detection using CNN activations. In: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017 (2017) 2, 3, 7, 8
11. Leung, T.K., Malik, J.: Detecting, localizing and grouping repeated scene elements from an image. In: Computer Vision - ECCV'96, 4th European Conference on Computer Vision, Proceedings, Volume I. Springer (1996) 2
12. Li, F., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. (2007) 8
13. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. In: IEEE international conference on computer vision ICCV (2019) 1

14. Liu, J., Liu, Y.: GRASP recurring patterns from a single view. In: IEEE Conference on Computer Vision and Pattern Recognition (2013) 2, 3, 7, 8
15. Liu, Y., Collins, R.T., Tsin, Y.: A computational model for periodic pattern perception based on frieze and wallpaper groups. IEEE Trans. Pattern Anal. Mach. Intell. (2004) 3
16. Liu, Y., Cheng, M.M., Hu, X., Bian, J.W., Zhang, L., Bai, X., Tang, J.: Richer convolutional features for edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 3
17. Logothetis, N.K., Sheinberg, D.L.: Visual object recognition. Annual review of neuroscience (1996) 1
18. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR (2014) 1
19. Ono, Y., Trulls, E., Fua, P., Yi, K.M.: Lf-net: Learning local features from images. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, NeurIPS (2018) 3
20. Park, M., Brocklehurst, K., Collins, R.T., Liu, Y.: Deformed lattice detection in real-world images using mean-shift belief propagation. IEEE Trans. Pattern Anal. Mach. Intell. (2009) 2
21. Pritts, J., Chum, O., Matas, J.: Rectification, and segmentation of coplanar repeated patterns. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2014) 2, 3
22. Rodríguez-Pardo, C., Suja, S., Pascual, D., Lopez-Moreno, J., Garces, E.: Automatic extraction and synthesis of regular repeatable patterns. Comput. Graph. (2019) 2, 3
23. Schaffalitzky, F., Zisserman, A.: Geometric grouping of repeated elements within images. In: Carter, J.N., Nixon, M.S. (eds.) Proceedings of the British Machine Vision Conference 1998, BMVC 2
24. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007). IEEE Computer Society (2007) 3
25. Singh, B., Davis, L.S.: An analysis of scale invariance in object detection snip. In: IEEE conference on computer vision and pattern recognition CVPR. pp. 3578–3587 (2018) 1
26. Stutz, D., Hermans, A., Leibe, B.: Superpixels: An evaluation of the state-of-the-art. Comput. Vis. Image Underst. (2018) 5
27. Torii, A., Sivic, J., Okutomi, M., Pajdla, T.: Visual place recognition with repetitive structures. IEEE Trans. Pattern Anal. Mach. Intell. (2015) 2
28. Wang, H., Zhao, G., Yuan, J.: Visual pattern discovery in image and video data: a brief survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. (2014) 3
29. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: IEEE conference on computer vision and pattern recognition CVPR (2017) 1