Egocentric Hierarchical Visual Semantics

Luca ERCULIANI [a], Andrea BONTEMPELLI [a,1], Andrea PASSERINI [a], and Fausto GIUNCHIGLIA [a]
[a] *University of Trento*

Abstract. We are interested in aligning how people think about objects and what machines perceive, meaning by this the fact that object recognition, as performed by a machine, should follow a process which resembles that followed by humans when thinking of an object associated with a certain concept. The ultimate goal is to build systems which can meaningfully interact with their users, describing what they perceive *in the users' own terms*. As from the field of *Lexical Semantics*, humans organize the meaning of words in hierarchies where the meaning of, e.g., a noun, is defined in terms of the meaning of a more general noun, its *genus*, and of one or more differentiating properties, its *differentia*. The main tenet of this paper is that object recognition should implement a hierarchical process which follows the hierarchical semantic structure used to define the meaning of words. We achieve this goal by implementing an algorithm which, for any object, recursively recognizes its *visual genus* and its *visual differentia*. In other words, the recognition of an object is decomposed in a sequence of steps where the locally relevant visual features are recognized. This paper presents the algorithm and a first evaluation.

Keywords. Genus and Differentia, visual semantics, interactive machine learning

1. Introduction

Lexical Semantics studies how word meanings, i.e., linguistic concepts [1,2] are formed, where these concepts are assumed to be constructed by humans through language. As from this field, humans organize the meaning of words in hierarchies where the meaning of, e.g., a noun, is defined in terms of a more general noun, its `Genus`, and of one or more differentiating properties, its `Differentia`. Thus for instance, a *guitar* is a *stringed (musical) instrument* with *six strings* [3]. The main tenet of the work described in this paper is that object recognition should implement a process which progressively visually reconstructs the hierarchical semantic structure used to define the meaning of words. Only in this way it is possible to have a full one-to-one *alignment* between how people think of the world and, ultimately, human language, and machine perception. The ultimate goal is to build systems which can meaningfully interact with their users, describing what they perceive *in the users' own terms*. Notice how this is a well known, still unsolved problem, the so called *Semantic Gap problem*, which was identified in 2010 [4] as (quote) "*... the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.*".

Based on the work in the field of *Teleosemantics* [5], see in particular the work in [6,7,8,9], the field of *Visual Semantics* has been introduced as the study of how hu-

[1] Corresponding Author: Andrea Bontempelli, andrea.bontempelli@unitn.it

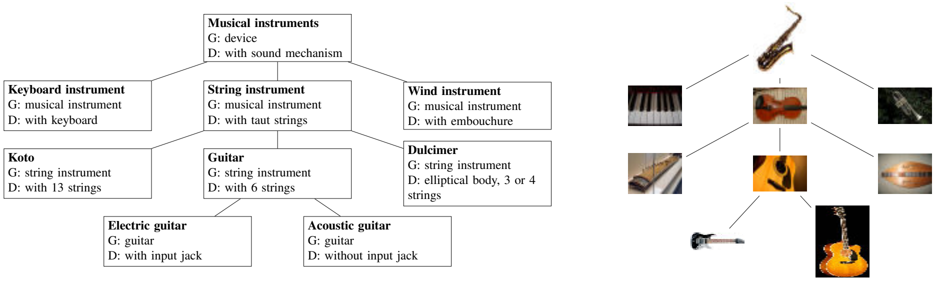

Figure 1. A classification concept hierarchy for Musical Instruments [13]. **Left**: *Lexical semantic* hierarchy and `Genus` (G) and `Differentia` (D) of each concept. **Right**: *Visual semantic* hierarchy.

mans build concepts when using vision to perceive objects in the world [10]. According to this line of work, objects should be recognized by recognizing first their genus and then their differentia, as visually represented in the input images or videos. Thus, for instance, a *guitar* should be recognized first as a *stringed instrument* (`Genus`), which is itself a *musical instrument* with *strings*, and then by recognizing its *six strings* (`Differentia`) [3,2]. This clearly leads to a recursive recognition process where the set of possible objects gets progressively restricted to satisfy more and more refined differentiae. In the most general case, the root node is the concept *object* itself, namely anything that can be detected as such, e.g., via a bounding box. In this context, we adopt an egocentric point-of-view with respect to a specific person [11,12].

As an example consider Fig. 1, taken from a small classification of musical instruments [13]. In this figure (left), we can see how the meaning of each label is provided in terms of a genus and a differentia, and where the genus one level down is the label of the concept the level up. Dually in the figure (right), we can see how all the images clearly show the differentia that allows to differentiate the object one level down from the object one level up (as having an extra feature, i.e., the visual differentia) and from all the siblings (as all objects under the same visual genus have a different visual differentia). Notice how, in current hierarchical computer vision tasks, the hierarchy is usually a-priori and static, e.g., [14], and does not consider the users' language and its mapping to their visual perception (see, e.g., [15,16]), leading therefore to a human-machine misalignment. For instance, a non-expert user would correctly classify the Koto instrument in Fig. 1 as a stringed instrument but, differently from a musician, would not describe it using its name, thus having two different but consistent linguistic descriptions of the same image.

The main goal of this paper presents a general algorithm which aligns machine perception and human description. The algorithm is based on two key ideas:

- Object recognition is implemented following a *hierarchical decomposition* process where the uniquely identifying features of the input object (the differentia) are recognized following the same order that is used in constructing the meaning of the label naming the object.
- Object recognition follows an *egocentric, incremental approach* where the user progressively refines the level of detail at which an object is recognized.

The work in [10] introduced visual semantics in the base hierarchy-less case. This paper extends this work to hierarchies of any depth. We do this by leveraging *Extreme Value*

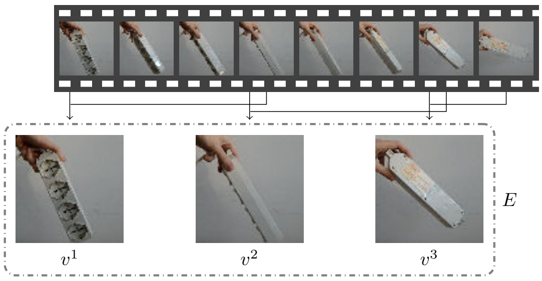

Figure 2. Example of an encounter. The video contains eight frames of a power strip gradually rotated over time on a white background. Similar adjacent frames are aggregated in three visual objects, which form the encounter E [10]. For better visualization, each visual object is represented, here and below, as its first frame.

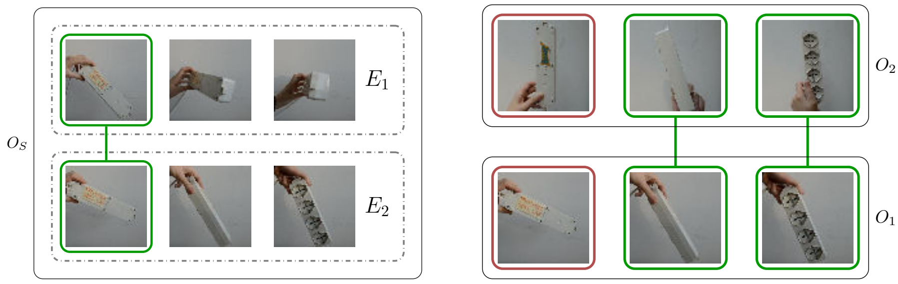

Figure 3. Left: An object made of two distinct encounters (dashed line), with their similar visual objects connected in green. **Right**: Two distinct objects sharing the same genus (via the visual objects connected in green). Red visual objects are the differentia (i.e., different tape on the back side).

Machines [17], a principled approach to open set problems which allows to implement differentia-based object recognition. The source code of the algorithm, the dataset and all the material necessary to reproduce the experiments are freely available online.[2]

2. Visual Semantics

We inherit from [10] the following foundational notions. An *encounter* E is an event during which a user sees an object. We computationally model an encounter as one or more *visual objects*, where a visual object consists of a *sequence of adjacent frames* that are similar to each other. Fig. 2 shows an example of an encounter with its decomposition into visual objects. An *object* O is a collection of encounters that are perceived to represent the same concept.

The left part of Fig. 3 shows an object consisting of two encounters, with the visual objects that make the two encounters similar highlighted in green. Two encounters that have at least a pair of visual objects that are similar are said to share the same `Genus`. Two encounters with the same `Genus` could or could not be associated with the same object. What makes this decision is the presence (or the absence) of a `Differentia`, i.e., a pair of visual objects that identifies the two encounters as representing two distinct objects. The right part of Fig. 3 visually presents these concepts. The visual objects con-

[2] `https://github.com/lucaerculiani/hierarchical-objects-learning`.

Algorithm 1 The main loop of the framework.

```
procedure MAIN
    while True do
        E ← PERCEIVE()
        O_e ← PREDICTGENUS(E)
        while not GENUSOF(E, O_e) do
            O_e ← PARENT(O_e)
        REFINEGENUS(O_e, E)
```

nected in green indicate that the two objects share the same `Genus`, while those circled in red are their `Differentia`, and indicate that they are distinct objects. The intuition is that some partial views of the objects determine their `Genus` and `Differentia` respectively. The hierarchy $\mathcal{H}$ organizes the objects, modeled as visual objects, in a tree-like structure, which outlines the subsumption relationships between the objects in terms of `Genus` and `Differentia` (see, e.g., Fig. 1).

3. Building an Egocentric Visual Semantic Hierarchy

The visual hierarchy is built incrementally as the objects are perceived. The interaction with the user ensures that the visual hierarchy matches the user lexical semantics. The proposed framework consists of a cyclic procedure in which at each iteration a new encounter (a sequence depicting an object) is shown to the model. The model then asks the user a series of queries over the `Genus` and `Differentia` of the new encounter with respect to some of the objects that were seen in the past by the algorithm (which are stored in its internal memory). Via this interaction, the user can guide the algorithm to assign the new encounter to the correct position inside the machine's knowledge base.

The main loop. Algorithm 1 lists the pseudo-code of the infinite learning loop that takes as input a new sequence at each iteration. This new sequence is first forwarded to an embedding algorithm that converts the video sequence, currently encoded as a series of frames, into a collection of visual objects (i.e. the encounter E). In this step, we employ an unsupervised deep neural network, pre-trained on a self-supervised class-agnostic task [18]. This training approach ensures that the embeddings are not explicitly biased towards the classes. Then, the PREDICTGENUS procedure searches in its memory for the most specific `Genus` O_e for encounter E, driven by its similarity with previously encountered objects. Starting from O_e, it interacts with the user to find the right position of the encounter, possibly updating the hierarchy during the process. First, the user could say that O_e is not a `Genus` of E (meaning that their common `Genus` is further up in the hierarchy). If this is the case, the algorithm goes up through the hierarchy until it finds a valid `Genus` for the encounter. This is refined by further interacting with the user via the REFINEGENUS procedure. The two procedures are detailed below.

Genus prediction. The PREDICTGENUS procedure outlined in Algorithm 2 continuously performs open-world recognition over the evolving $\mathcal{H}$ of objects seen so far by the machine. The task is to predict the most specific node O_e in $\mathcal{H}$ of the current encounter E. The algorithm computes also the probability p_e that E belongs to O_e. For every visual object v of the current encounter, the candidate `Genus` is identified. Then, the al-

Algorithm 2 The procedure predicting `Genus` of an encounter.

function PREDICTGENUS(E)
  $O_e \leftarrow$ GETROOT($\mathcal{H}$)
  $p_e \leftarrow 1.0$
  for $v \in E$ **do**
    $O_v, p_v =$ PREDICTVOGENUS(v, GETROOT($\mathcal{H}$), 1.0)
    if $O_e =$ GETROOT($\mathcal{H}$) $\vee\, p_e < p_v$ **then**
      $O_e \leftarrow O_v$
      $p_e \leftarrow p_v$
  return $\langle O_e, p_e \rangle$

function PREDICTVOGENUS(v, O_v, p_v)
  $\lambda \leftarrow$ GETREJECTIONTRESHOLD($\mathcal{K}$)
  $\mathcal{C} \leftarrow$ GETCHILDREN(O_v)
  $p_{c^*} \leftarrow max_{\,O_c \in \mathcal{C}}$ PROBABILITY(O_c, v)
  $O_{c^*} \leftarrow arg\,max_{\,O_c \in \mathcal{C}}$ PROBABILITY(O_c, v)
  if $p_{c^*} > \lambda$ **then**
    return PREDICTVOGENUS(v, O_{c^*}, p_{c^*})
  else
    return $\langle O_v, p_v \rangle$

gorithm outputs the one with maximal probability (excluding the root node, which has always probability 1.0). The procedure PREDICTVOGENUS in Algorithm 2 computes the `Genus` of a visual object by navigating down the hierarchy. By leveraging the notion of `Genus` and `Differentia`, the algorithm searches for the most specific node that represents the current visual object. Starting from the root of the hierarchy, it computes the most probable genus child O_{c^*} for the visual object. If its probability p_{c^*} exceeds a rejection threshold λ, the procedure recurs over it, otherwise it stops returning its parent as the `Genus` of the visual object. Following [11], the rejection threshold is chosen with an optimization procedure that maximizes the number of correct predictions given the previous feedback by the user (stored in a supervision memory $\mathcal{K}$). See the original paper for the details. The procedure terminates when the most probable node is either below the threshold or is a leaf.

The probability of a `Genus` being the genus of a visual object can be formalized as the probability that an element belongs to a set (the set of previous visual objects for that `Genus`). To compute this inclusion probability, we employed the Extreme Value Machine (EVM) framework [17], because of its soundness and practical effectiveness. Basically each node is associated with a set of examples, called *extreme vectors*, that are used as representatives of the corresponding `Genus`. We use as extreme vectors all the visual objects associated with any of the nodes in the subtree of the node. The probability for a new visual object is computed based on the closest extreme vector and its associated probability distribution (a Weibull distribution).

Genus refinement. Starting from the most specific `Genus` for the encounter E that has been identified by the machine *and* confirmed by the user (called the *current* `Genus` and corresponding to ”thing” in the simplest case), the algorithm traverses the hierarchy down asking questions to the user to further refine the `Genus` for E. Fig. 4 presents a

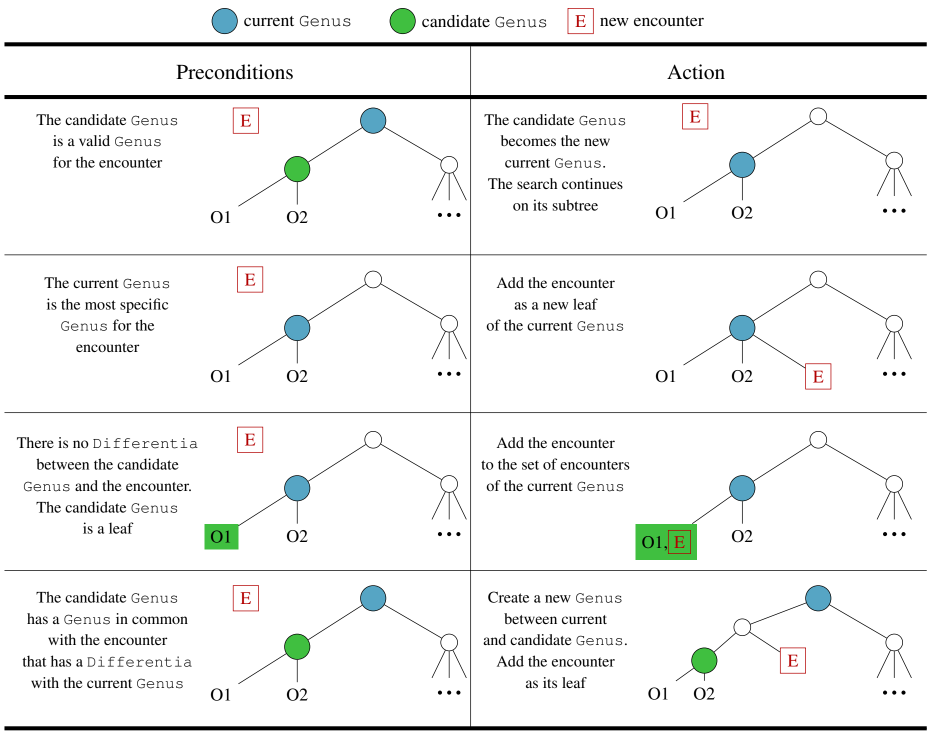

Figure 4. Representation of four possible choices that can be taken during the iterative encounter procedure.

graphical representation of the four possible situations at each iteration of the algorithm, and the action to be taken as a result of the user feedback. The new encounter is depicted in red, the current `Genus` (current best guess for `Genus`) in cyan, while the green node is the `Genus` for which the machine is asking queries. For each of the possible actions, each row in the table shows the preconditions that must be met in terms of `Genus` and `Differentia` between the encounter, the candidate `Genus` and the current `Genus`, and the effect that each action has on the inner hierarchy of the machine. The REFINE-GENUS procedure consists of a sequence of questions and corresponding actions, as reported in Fig. 4, until one of the actions results in placing the new encounter E in the hierarchy (one of the two lower actions in the figure).

4. Experiments

The goal is to evaluate how much the hierarchy built by the machine is aligned with the user hierarchy. This evaluation is done by measuring the distance in the hierarchy between the predicted `Genus` and the user desired `Genus`. The greater the distance, the greater the misalignment. All experiments were implemented in Python3 and PyTorch.

Data set. We used a collection of objects organized in a perfectly balanced hierarchy of 4 levels, such that each node (except for the leaves) inside the hierarchy has 3 children, leading to a total of $3^4 = 81$ leaves. Each object was recorded 5 different times while

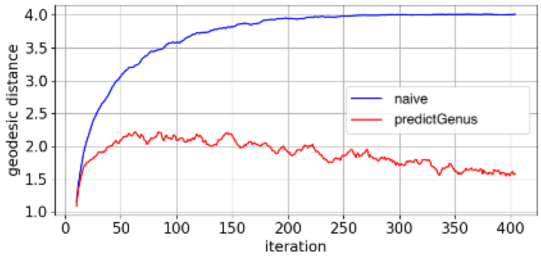

Figure 5. Comparison between *naive* model and PREDICTGENUS in terms of geodesic distances between the predicted and the correct `Genus`.

rotated or deformed against a uniform background, thus obtaining 405 encounters. The hierarchy was used to simulate the supervision of the user.

Experimental details. The whole set of videos in the dataset is presented to the machine in random order. An agent simulates the user and provides supervision to the model by comparing the ground-truth hierarchy of the data set against the hierarchy of the machine, and replying to the queries of the algorithm accordingly (see Fig. 4). The hierarchy of objects that is built over time is always consistent with the ground truth. The machine goal is to minimize the categorization effort required from the user when a new encounter must be placed into the hierarchy. The machine suggests the starting node, from which the user navigates down the hierarchy until the correct node is found. The closer the prediction of the machine is to the ground-truth node, the lower the effort of the user.

The performance is evaluated in terms of *geodesic distance*, namely the number of edges in the shortest path between the predicted node and the target node selected by the user. Even if this measure is affected by the size of the hierarchy (the deeper the tree, the greater the average distance between couples of nodes), due to the fact that the evolution of the hierarchy is completely guided by the user, any model has always its hierarchy updated in the same way. This fact keeps this performance measure unbiased in the context of this experiment. We compare the performance PREDICTGENUS with that of a *naive* algorithm that always suggests the root of $\mathcal{H}$ as the starting node.

Results. Fig. 5 reports the geodesic distance when varying the number of iterations of the algorithm, averaged over 100 runs with different random orderings of the objects. Our model, shown in red, substantially outperforms the naive algorithms (in blue), with the difference becoming more pronounced as the number of observed objects increases. After an initial phase in which the cost increases due to the rapid expansion of the hierarchy, the average geodesic distance for PREDICTGENUS starts to decrease (at roughly 60 iterations). The algorithm suggests starting nodes closer to the correct node with higher accuracy because the increasing amount of encounters allows to better model each `Genus`. In contrast, the naive algorithm converges toward an average cost that is equal to the average distance between the root node and the leaves (the hierarchy of the dataset is composed by four levels). Notice that a particularly bad predictor could in principle do worse than the naive algorithm, by predicting a node in a subtree that is more distant than the root node. Albeit preliminary, these results confirm the potential of the proposed framework in correctly acquiring the hierarchy of the user and its semantics in terms of `Genus` and `Differentia`.

5. Related work

Our work implements an egocentric, incremental object recognition. A closely related area is continual learning, which addresses the problem of learning to recognize novel objects without forgetting previous knowledge [19]. Traditional machine learning assumes a closed-world setting, where the set of classes is defined at training time. Open set approaches reject examples belonging to *unknown unknown* classes [20]. Studies on open world recognition extend open set by incrementally updating the model to incorporate the new classes [21,22,23].

In many real-world tasks, such as gene classification and music genre recognition [24], the target labels have hierarchical relationships. In computer vision, [15,16] tackle hierarchical novelty detection by identifying to which node in the hierarchy the novel class is attached. Hierarchical information has been used to achieve more reasonable classification errors [25,14] or integrated into neural networks [26,27]. These works differ in that do not consider the semantic and egocentric aspects.

In the works on visual-semantic embeddings, the idea is to map the input feature space to a semantic embedding space [28,29], for instance by projecting the images and the knowledge graph into a unified representation [30]. [31] learns object attributes, both semantic (part of the objects) and non-semantic (visual feature space), from annotations to classify images. These approaches differ in that they neither try to align recognition with lexical semantics nor use hierarchical classifications.

The approaches that study the grounding of human language in perception, especially vision [32], are strongly related to our work. Examples in this field are answering questions grounded on visual images [33], image captioning [34], visual commonsense [35] and visual reasoning with natural language [36]. These approaches do not leverage the work done in lexical semantics to drive the object recognition process.

6. Conclusion

In this paper, we have introduced a novel approach where objects are recognized following the same hierarchical process that is used in lexical semantics to provide meaning to the nouns used to name objects. The first set of experiments shows a consistent improvement in the alignment between what the system recognizes and the words used by humans to describe what is being recognized.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 FET Proactive project "WeNet - The Internet of us", grant agreement No. 823783, and from "DELPhi - DiscovEring Life Patterns" project funded by the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2017 – DD n. 1062 del 31.05.2019. The research of AP was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] Kuang Z, Yu J, Li Z, Zhang B, Fan J. Integrating multi-level deep learning and concept ontology for large-scale visual recognition. Pattern Recognition. 2018;78:198-214.

[2] Giunchiglia F, Batsuren K, Bella G. Understanding and Exploiting Language Diversity. In: IJCAI; 2017. p. 4009-17.

[3] Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ. Introduction to WordNet: An on-line lexical database. International journal of lexicography. 1990;3(4):235-44.

[4] Smeulders A, Worring M, Santini S, Jain R. Content-based image retrieval at the end of the early years. IEEE Transactions on PAMI. 2000;22(12):1349-80.

[5] Macdonald G, Papineau D, et al. Teleosemantics. Oxford University Press; 2006.

[6] Millikan RG. Language, thought, and other biological categories: New foundations for realism. MIT press; 1984.

[7] Millikan RG. A more plausible kind of "recognitional concept". Philosophical Issues. 1998;9:35-41.

[8] Millikan RG. On clear and confused ideas: An essay about substance concepts. Cambridge University Press; 2000.

[9] Millikan RG. Language: A biological model. Oxford University Press on Demand; 2005.

[10] Giunchiglia F, Erculiani L, Passerini A. Towards visual semantics. SN Computer Science. 2021;2(6):1-17.

[11] Erculiani L, Giunchiglia F, Passerini A. Continual egocentric object recognition. ECAI. 2020.

[12] Smith LB, Slone LK. A Developmental Approach to Machine Learning? Frontiers in Psychology. 2017;8.

[13] Giunchiglia F, Bagchi M, Diao X. Aligning Visual and Lexical Semantics. In: 18th International Conference on Information (iConference). Springer; 2023. Available from: `https://arxiv.org/abs/2212.06629`.

[14] Bertinetto L, Mueller R, Tertikas K, Samangooei S, Lord NA. Making better mistakes: Leveraging class hierarchies with deep networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 12506-15.

[15] Lee K, Lee K, Min K, Zhang Y, Shin J, Lee H. Hierarchical novelty detection for visual object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. .

[16] Ruiz I, Serrat J. Hierarchical novelty detection for traffic sign recognition. Sensors. 2022.

[17] Rudd EM, Jain LP, Scheirer WJ, Boult TE. The extreme value machine. TPAMI. 2018.

[18] Caron M, Bojanowski P, Mairal J, Joulin A. Unsupervised pre-training of image features on non-curated data. In: ICCV; 2019. .

[19] De Lange M, Aljundi R, Masana M, Parisot S, Jia X, Leonardis A, et al. A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence. 2021.

[20] Scheirer WJ, de Rezende Rocha A, Sapkota A, Boult TE. Toward open set recognition. IEEE transactions on pattern analysis and machine intelligence. 2012.

[21] Bendale A, Boult TE. Towards open world recognition. In: CVPR; 2015. .

[22] De Rosa R, Mensink T, Caputo B. Online open world recognition. arXiv preprint arXiv:160402275. 2016.

[23] Geng C, Huang Sj, Chen S. Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence. 2020.

[24] Simeone P, Santos-Rodríguez R, McVicar M, Lijffijt J, De Bie T. Hierarchical novelty detection. In: Advances in Intelligent Data Analysis XVI: 16th International Symposium, IDA 2017, London, UK, October 26–28, 2017, Proceedings 16. Springer; 2017. .

[25] Deng J, Krause J, Berg AC, Fei-Fei L. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012. .

[26] Deng J, Ding N, Jia Y, Frome A, Murphy K, Bengio S, et al. Large-scale object classification using label relation graphs. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer; 2014. .

[27] Nauta M, Van Bree R, Seifert C. Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 14933-43.

[28] Fu Y, Dong H, Ma Yf, Zhang Z, Xue X. Vocabulary-informed extreme value learning. arXiv preprint arXiv:170509887. 2017.

[29] Zhao H, Puig X, Zhou B, Fidler S, Torralba A. Open vocabulary scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. .

[30] Lonij V, Rawat A, Nicolae MI. Open-world visual recognition using knowledge graphs. arXiv preprint arXiv:170808310. 2017.

[31] Farhadi A, Endres I, Hoiem D, Forsyth D. Describing objects by their attributes. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009. .

[32] Baroni M. Grounding distributional semantics in the visual world. Language and Linguistics Compass. 2016.

[33] Bernardi R, Pezzelle S. Linguistic issues behind visual question answering. Language and Linguistics Compass. 2021.

[34] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research. 2013.

[35] Zellers R, Bisk Y, Farhadi A, Choi Y. From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. .

[36] Suhr A, Zhou S, Zhang A, Zhang I, Bai H, Artzi Y. A Corpus for Reasoning about Natural Language Grounded in Photographs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2019. .