Ontology Based Image Recognition: A Review

Sandeepak Bhandari
Department of Software Engineering
Kaunas University of Technology
Kaunas, Lithuania
Sandeepak525@gmail.com

Audrius Kulikajevas
Department of Multimedia Engineering
Kaunas University of Technology
Kaunas, Lithuania
akulikajevas@gmail.com

***Abstract* — Due to lack of domain knowledge about the semantics of the image, image retrieval rate is usually unsatisfying. To improve this, image labels must be provided by the author of the dataset. Applying ontology in digital image recognition to extract relevant information such as timeline, features, visualization to help understand and interpret it, so that we can focus on the most relevant information.**

Keywords — Ontology, features, image labeling, classification, semantic gap, cognitive vision.

I. Introduction

Computer vision as of late has seen a resurgence in popularity with the increased interests in machine learning, specifically neural networks. Despite of this, majority of work is restricted to specific domain knowledge such as specific object instance recognition, face recognition, etc. which makes these type of computer systems lacks the flexibility and adaptability to different domain object recognition. The term cognitive vision [1] has been introduced to encapsulate an attempt to achieve more robust and adaptive computer vision systems. Cognitive vision systems can infer spatial relations such as an ocean is often near a beach, a lake is besides a grassland [2]. With this paper we try to evaluate previous research done on object recognition tasks with the addition of semantics in a form of domain ontologies. This paper is organized as follows: Section II provides short introduction in related work. Section **Errore. L'origine riferimento non è stata trovata.** provides in-depth analysis of Semantic Web applications in the field of object recognition and tasks related to it, such as segmentation. Section IV describes possible applications of the Semantic Web in the domain specific object recognition tasks. Finally, Section **Errore. L'origine riferimento non è stata trovata.** provides our conclusions to the effectiveness of using ontology in the field of computer vision.

II. Related work

There has been a variety of recent studies in the field of cognitive computing. With the rise of machine learning and neural networks it is worth to re-evaluate the benefits of applying ontologies to existing object classification and recognition tasks. One of the most important tasks when it comes to object recognition is the creation of image labeling system for the identification of ground truths. Some of the possibilities to create the ground truths involve image annotation by keywords, free text annotations or annotations based on ontologies [3] which allows to add hierarchical structure to collection of keywords in order to produce a taxonomy. Furthermore, ontologies can be used for activity recognition from video feed [4] [5], allowing cognitive vision systems to semantically identify activities from low-level events. Traditional object recognition methods rely on detecting an individual object as a whole, ontology based methods are capable of detecting objects based on their individual component compositions [6]. Thus, making ontology a powerful tool to be used in the future of computer vision and machine learning.

Anne-Marrie Tousch et al. [7] have made contributions in survey of semantic techniques used for image annotation. In the paper, authors have made an analysis about the nature of semantics to describe the image, where they pointed out three main levels to describe the semantics. First level describes objective and concrete object descriptions in relation to abstract ones for example a crying person is an objective description, while inferred pain would be subjective one without deeper knowledge of the semantic context. Second level compares generic versus specific objects, also referred to as individual instances in object recognition, i.e. a bridge vs. Golden gate bridge. Third and final semantics level is split into four facets of time, localization, events or objects. Another contribution to the discussion was regarding the semantic analysis, explanation of the semantic gap and observations on discrepancies between human ability to recognize almost instantly an enormous number of objects versus the possibilities of image recognitions done by machines. They clarify the term semantics in the context that they (and by extension us) use in their paper as an image description in natural language, with semantic analysis referring to any kind of transcription of an image into linguistic expression. Furthermore, they describe several approaches for semantic analysis using unstructured vocabularies such as:

1. Direct methods using plain representation of data and plain statistical methods.
2. Linguistic methods based on the use of an intermediate visual vocabulary between raw numerical data and high-level semantics.
3. Compositional methods where parts of an image are identified before the whole image or its parts are annotated.
4. Structural methods where a geometry of parts is used.

5. Hierarchical compositional methods where a hierarchy of parts is constructed for recognition.
6. Communicating methods when information is shared between categories.
7. Hierarchical methods that search for hierarchical relationships between categories.
8. Multilabel methods assigning several global labels simultaneously to an image.

Finally, authors have touched on semantic image analysis using structured vocabularies, where they distinguish two main relationship types commonly found in ontologies: Is-A-Part-Of and Is-A relationship. Where the later signifies inheritance and former signifies composition, such as being part of a bigger model. However, according to authors such relationships are not descriptive enough and suggest finer organizations of methods on how semantic relations are introduced in the system such as:

1. Linguistic methods where the semantic structure is used at the level of vocabulary, independently from the image, e.g. to expand the vocabulary.
2. Compositional methods that use meronyms, e.g. components.
3. Communication methods that use semantic relations to share information between concepts, be it for the feature extraction or for classification.
4. Hierarchical methods use Is-A relations to improve categorization and/or to allow classification at different semantic levels.

There have been other contributions to the discussion of automatic image annotation techniques by applying semantics [8] where author review on different approaches of automatic image annotation are reviewed: 1) generative model based image annotation, 2) discriminative model based image annotation, 3) graph model based image annotation. Due to the rapid advancement of digital technology in the last few years, there has been an increasingly large number of images available on the Web, making manual annotation of images an impossible task. However, we will be focusing on the steps of applying semantics to each step of image recognition individually along with applications of semantics to a narrow domain.

III. Semantic web in object recognition

Image retrieval is the key task to be solved in the science of computer vision. While more classical approaches had to an extent worked in the past for simple object recognition, more general approaches are required in the modern times. In this section we present research done in the past on application of ontology to some steps of image recognition to improve image retrieval performance. We also present TABLE I. for comparison between such methods.

A. Semantics application in image segmentation

One of the main tasks in any image recognition software is the image segmentation step. During this step, an image is segmented into viable detection Regions of Interest (ROIs) to optimize the following recognition steps. However, image segmentation is a very challenging, albeit necessary task for any kind of image recognition algorithm therefore any optimizations of it are a welcome addition to any cognitive vision system. In this section of the paper we will focus on application of the semantic web technologies in order to optimize the results of the segmentation algorithms. The quality of recognition highly depends on the ability to segment any given frame. Previously classical algorithms such as watershed have been used to achieve image segmentation. However, such algorithms had lacked the efficiency, precision and robustness in real world scenarios where occlusion, motion and illumination play key roles in the scene. There has been some research in adapting convolutional neural networks for the segmentation task such in techniques such as Mask R-CNN [9]. However, due to the fact that CNNs highly depend on the domain they have been trained on, they have troubles detecting changes even in the domains that they are familiar. In the paper [10] a method was proposed which involves simultaneous image segmentation and detection of simple objects, imitating partially how the human vision works. The initial region labeling is performed based on regions low-level descriptors with concepts stored in an ontological knowledge base. This allows the proposed technique to associate each region to a fuzzy set of candidate labels. Afterwards, a merging process is performed based on new similarity measurements and merging criteria that are defined at the semantic level with the use of fuzzy set operations. Furthermore, this approach is invariant to the chosen region growing algorithm and can be applied to any of them with certain modifications, to demonstrate this, authors apply semantics to watershed and RSST segmentation, experimentally showing that semantic watershed had 90% accuracy, semantic RSST accuracy of 88% compared to classical RSST approach of 82%. Other experiments have shown a similar increase in accuracy of 7-8%. To achieve these improvements, authors have proposed to adjust merging processes as well as termination criteria of the classical region growing algorithms. What is more, a novel ontological representation for context is introduced, combining fuzzy theory and fuzzy algebra with characteristics derived from the semantic web, such as reification. Membership degrees of labels are assigned to regions derived from the semantic segmentation are re-estimated appropriately, according to context-based membership degree readjustment algorithm, which utilizes ontological knowledge, to optimize membership degrees for the detected concepts of each region in the scene. While the contextualization and initial region labeling steps are domain specific and require the domain of the images to be provided as input, the rest of the approach is domain independent. In another paper [11], authors have shown that applying semantics to a convolutional neural network for the task of image segmentation can greatly increase performance. Authors have experimentally proven that a neural network that is trained end-to-end, pixel-to-pixel on semantic segmentation can achieve the best asymptotical and absolute performance results without the downside of other methods such as patch-wise training that lack the efficiency of convolutional training, or the needed inclusion of superpixels such as the ones used in [12] where they are used to generate semantic objects parts. However, the superpixels used in the later additionally give a bridge between low-level and high-level features by

incorporating semantic knowledge allowing to infer labels of individual segmented regions.

B. Semantics application to image labeling

Convolutional neural networks have shown great performance in their ability to correctly detect the objects and actions in the domains that they are familiar with. However, in order for the network to be able to interpret the image it sees correctly it needs a vast amount of data samples from which to compare against. There already exists databases of labeled images such as ImageNet that can provide useful training data, although with the rapid development of social media, automatic techniques capable of effectively understanding and labeling the media are required. Content aware systems are capable of indexing, searching, retrieving, filtering and recommending multimedia from the vast quantity of media posted of social media. However, such unconstrained data has very high complexity of objects, events and interactions in the consumer videos. Such unconstrained domains create numerous problems for previously available video analysis algorithms, such as the ones capable of recognizing human activity. What is more, home videos being the most prevalent genre suffer from increased good feature extraction problems due to poor lightning conditions, occlusion, clutter in the scene, shaking and/or low-resolution cameras and various other background noise. In the reviewed paper [13] authors try to address these problems by introducing a new attribute-learning framework that learns a unified semilatent attribute space. Latent attributes are used to represent all shared aspects of the data, which are not explicitly included in users' sparse and incomplete annotations. `Latent attributes are used as complimentary annotations for user specified attributes and are instead discovered by the model through joint learning of semilatent attribute space. This gives authors a mechanism for semantic feature reduction from the raw data in multiple modalities to a unified lower dimensional semantic attribute space. These semilatent attributes are used to bridge the semantic gap with reduced dependence on completeness of attribute ontology and accuracy of the training attribute labels. Described method has given the authors the flexibility to learn a full semantic-attribute space of the video feed irrespective of how well defined and complete the user given data about it is. Furthermore, they have managed to improve multitask and N-shot learning by leveraging latent attributes, went beyond existing zero-shot learning approaches by exploiting latent attributes, leveraged attributes in conjunction with multimodal data to improve cross-media understanding, enabling new tasks such as explicitly learning which modalities attributes appear in. Finally, the proposed method is applicable to large multimedia data sets as it is expressed in a significantly more scalable way than previously available techniques, making the technique invariant to the length of the given input video or the density of available good features in it.

Other research [14] supports the notion that that multimedia resources "in the wild" are growing at a staggering rate and that the rapidly increasing number of multimedia resources has brought an urgent need to develop intelligent methods to organize and process them. In this paper, the Semantic Link Network model is used for organizing multimedia resources. Semantic Link Network (SLN) is designed to establish associated relations among various resources (e.g., Web pages or documents in digital library) aiming at extending the loosely connected network of no semantics (e.g., the Web) to an association-rich network. Since the theory of cognitive science considers that the associated relations can make one resource more comprehensive to users, the motivation of SLN is to organize the associated resources loosely distributed in the Web for effectively supporting the Web intelligent activities such as browsing, knowledge discovery and publishing, etc. The tags and surrounding texts of multimedia resources are used to represent the semantic content. The relatedness between tags and surrounding texts are implemented in the semantic Link Network model. The data sets including about 100 thousand images with social tags from Flickr are used to evaluate the proposed method. Two datamining tasks including clustering and searching are performed by the proposed framework, which shows the effectiveness and robust of the proposed framework.

C. Semantic web application to recognize the object based on individual parts

Semantic web (ontologies) gives us the powerful ability to infer certain attributes about the object based on the domain knowledge of the individual object components of said object. Allowing us to instead of recognizing a specific object instance or it's class to recognize individual components available in the scene and based on the known Is-A/Is-Part-Of relationships infer what kind of objects can be created with those components. This novel way of applying ontologies to the task of object recognition was the application of object semantics based on what the object is being used for [15]. In the case of the paper, the semantics were used for tool recognition where the type of tool is inferred by what it's functionality in relation to human hand is. In the work, authors assert that objects do not change functionality based on changes to their details such as a cup having multiple handles would still be considered a cup. Instead, they focus on object parts and their combinations to assign a function to a tool. This approach gives the advantage that the system is not trained to wantonly different individual objects and will instead detect parts that contribute to the fundamental tool functionality. The proposed method consists of three main stages: preprocessing, object signature extraction and object similarity calculations. Object signature extraction is subdivided into two steps: part signature extraction and pose signature extraction. During part signature extraction a support vector machine (SVM) is used to find the characteristic descriptions of given object. Pose signatures describe how parts are attached to each other which provides the information on how the parts are rotated to respect to one another and locations at which parts are connected to one another. Once the features are extracted function analyzer algorithm is ran which allows to compare objects and assign functional meanings to them thus, achieving these tasks: recognizing the object, generalize the object between different objects with different number of parts, assigning multiple functions to the same object, providing the ability to find another use for an object. Authors have shown that their method unlike deep convolutional neural networks do not require such extensive training sets and can generalize on very few training samples.

TABLE I. ONTOLOGY BASED IMAGE RETRIEVAL METHODS

S. No	Topics/Concepts	Pros	Cons
1.	Keyword based image retrieval (text based, field based, structure based)	Utilizations diverse watchwords; Use at least one picture properties; Keywords portraying picture data	Can't depict a picture totally and semantically
2.	Low-level feature: color (histogram and moments, dominant color, color cluster, etc.)	Shading likeness based recovery; Color cognizance vector based recovery	Shading alone can't depict the full picture content
3.	Low-level feature: shape (fourier transform, curvature scale, template tatching, etc.)	Template matching method	Shape alone can't depict the full picture content
4.	Low-level feature: texture (wavelet transform, edge statistics, Gabor filters, statistical based, etc.)	Shading likeness based recovery; Color cognizance vector based recovery	Shading and surface alone can't depict the full picture content
5.	Scale Invariant Feature Transform: SIFT	Foreseeing amino corrosive changes in Protein structure; Features for picture recovery	Science application; Concluded to discover better descriptor
6.	Speeded Up Robust Feature: SURF	Novel scale and revolution invariant element depiction; CBIR visual consideration demonstrate	Not totally indented to endeavor semantic hole filling
7.	Ontology based image retrieval methods	Study: CBIR with abnormal state semantics Ontology based intellectual vision	Basic ontology with restricted visual highlights

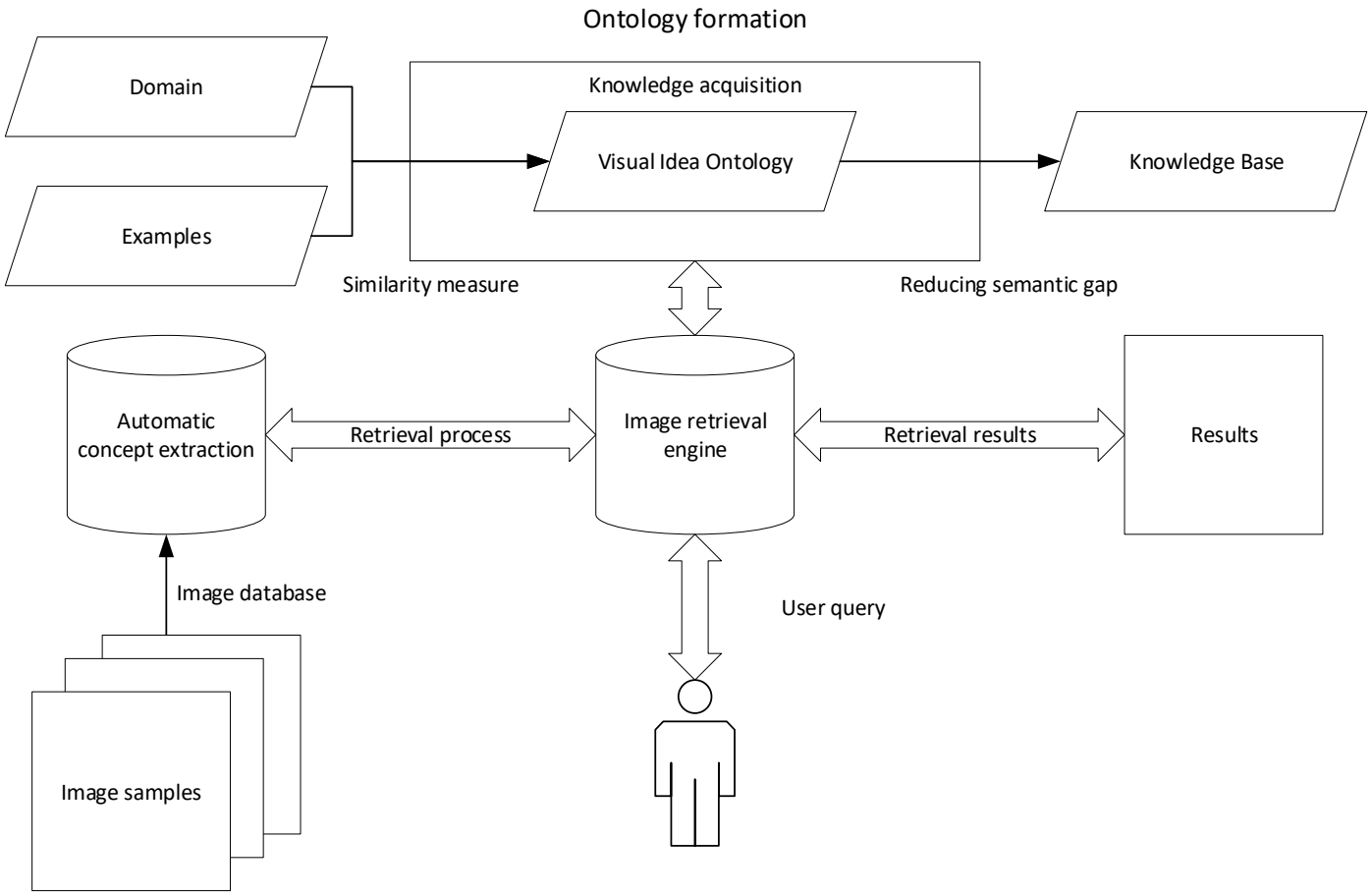

Fig. 1. Design ontology connected picture recovery process

The Fig. 1 setting portrayal could be surrounded utilizing metaphysics from the above said picture ideas. Applying Description Logics (DL), the information portrayal could be framed. The DAML (DARPA Markup Language) and OIL

(Ontology Interface Language) are utilized for this usage which is accessible with OWL (Web Ontology Language). Standards for portraying connection between picture includes in metaphysics can be characterized utilizing the DL moreover. Once the idea cosmology encircled (for instance spatial philosophy), the similitude coordinating of client inquiry with extricated picture highlight is evaluated through the metaphysics chain of command. This furnishes more closeness to client question with pictures in database. There are a few instruments have been produced to be specific "OntoVis" which perform three undertakings space learning securing, metaphysics driven visual securing and picture case administration. The advantage of utilizing visual thought metaphysics is to fill the semantic hole however much as could be expected amongst low and abnormal state idea

IV. SEMANTIC WEB IN DOMAIN SPECIFIC TASKS

A. Application in robotics

Mobile robots are one of the key applications that benefit from object recognition technologies. More and more robots are entering human living and working environments, where to operate successfully they are faced with multitude of real world challenges such as being able to handle many objects located in different places. To overcome these challenges a solution of robot efficiency needs to be found both in terms of computational efficiency and reducing the amount of detected false positives. Addition of ontology has shown to be beneficial in the object recognition tasks for methods such as *RoboEarth* [16] where the addition of such semantic mapping layer has experimentally shown to decrease computational time by only checking against the 10 most promising object annotations in large object databases. *RoboEarth* describes methodology they name as "action recipes" which semantically describe what action need to be performed to complete a specific task such as map the environment or to determine the location of an object. For example, "ObjectSearch" action recipe based on the prior partial room knowledge and its landmarks is capable to infer potential locations from where the desired object might be detected.

B. Application in Geographic Information Science

GEOBIA (Geographic Object-Based Image Analysis) [17] is not only a hot topic of current remote sensing and geographical research. It is believed to be a paradigm in remote sensing and Geographic Information Science (GIScience). It aims to develop automated methods for partitioning remote sensing (RS) imagery into meaningful image objects, and to assess their characteristics through spatial, spectral, textural, and temporal features, thus generating new geographic information in a GIS ready format.

Geographic Object-Based Image Analysis (GEOBIA) [18] represents the most innovative new trend for processing remote sensing images that has appeared during the last decade. However, its application is mainly based on expert knowledge, which consequently highlights important scientific issues with respect to the robustness of the methods applied in GEOBIA. In this paper, authors argue that GEOBIA would benefit from another technical enhancement involving knowledge representation techniques such as ontologies. Authors summarize the main applications of ontologies in GEOBIA, especially for data discovery, automatic image interpretation, data interoperability, workflow management and data publication. Among the broad spectrum of applications for ontologies, mention systems engineering, interoperability and communication. Because the method is a part of systems engineering, GEOBIA experts follow a series of analytical procedures to develop a system designed to produce geographic information. These procedures principally involve (i) data discovery and (ii) data processing and analysis, i.e., image interpretation.

C. Application to sports events

In this paper [19], authors present an ontology-based information extraction and retrieval system and its application in the soccer domain. In general, authors deal with three issues in semantic search, namely, usability, scalability and retrieval performance. Authors propose a keyword-based semantic retrieval approach. The performance of the system is improved considerably using domain-specific information extraction, inferencing and rules. Scalability is achieved by adapting a semantic indexing approach and representing the whole world as small independent models. The system is implemented using the state-of-the-art technologies in Semantic Web and its performance is evaluated against traditional systems as well as the query expansion methods. Furthermore, a detailed evaluation is provided to observe the performance gain due to domain-specific information extraction and inferencing.

Authors presented a novel semantic retrieval framework and its application in the soccer domain, which includes all the aspects of Semantic Web, namely, ontology development, information extraction, ontology population, inferencing, semantic rules, semantic indexing and retrieval. When these technologies are combined with the comfort of keyword based search interface, authors obtain a user-friendly, high performance and scalable semantic retrieval system. The evaluation results show that this approach can easily outperform both the traditional approach and the query expansion methods. Moreover, authors observed that the system can answer complex semantic queries without requiring formal queries such as SPARQL. Authors observe that the system can get close to the performance of SPARQL, which is the best that can be achieved with semantic querying. Finally, authors show how the structural ambiguities can be resolved easily using semantic indexing.

V. CONCLUSIONS

In this paper we present the importance and usefulness of applying ontology for image recognition tasks. We have reviewed different ontology based techniques, compared them to more classical approaches such as SIFT and SURF and provided with a list of benefits and possible drawbacks of using such techniques. With our paper we have concluded that applying semantics can greatly improve not only the overall performance of object recognition but also the performance and quality of individual tasks required for object recognition such as image segmentation. Moreover, we have found that ontology can be used to substantially reduce semantic gap i.e. the difference between the understanding of images by human and interpretation of images by machine, allowing for better automatization in training neural networks, as the dataset

preparation can be offloaded to a machine instead of being manufactured by hand. Finally, we have discussed the concept of semantic web, based on which the ontology can be formed.

REFERENCES

[1] P. Auer *et al.*, "A Research Roadmap of Cognitive Vision," *IST Proj. IST-2001-35454*, 2005.

[2] J. P. Schober, T. Hermes, and O. Herzog, "Content-based image retrieval by ontology-based object recognition," *KI-2004 Work. Appl. Descr. Logics*, no. January, 2004.

[3] A. Hanbury, "A survey of methods for image annotation," *J. Vis. Lang. Comput.*, vol. 19, no. 5, pp. 617–627, 2008.

[4] D. Tahmoush and C. Bonial, "Applying Attributes to Improve Human Activity Recognition," *Appl. Imag. Pattern Recognit. Work. (AIPR), 2015 IEEE*, 2015.

[5] U. Akdemir, P. Turaga, and R. Chellappa, "An ontology based approach for activity recognition from video," *Proceeding 16th ACM Int. Conf. Multimed.*, pp. 709–712, 2008.

[6] S. Tongphu, B. Suntisrivaraporn, B. Uyyanonvara, and M. N. Dailey, "Ontology-based object recognition of car sides," *2012 9th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2012*, 2012.

[7] A. M. Tousch, S. Herbin, and J. Y. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognit.*, vol. 45, no. 1, pp. 333–345, 2012.

[8] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognit.*, vol. 45, no. 1, pp. 346–362, 2012.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017.

[10] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias, "Semantic image segmentation and object labeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 298–311, 2007.

[11] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[12] M. Zand, S. Doraisamy, A. A. Halin, and M. R. Mustaffa, "Ontology-Based Semantic Image Segmentation Using Mixture Models and Multiple CRFs," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3233–3248, 2016.

[13] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Learning multimodal latent attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 303–316, 2014.

[14] U. Manzoor and M. A. Balubaid, "Semantic Image Retrieval : An Ontology Based Approach," vol. 4, no. 4, pp. 1–8, 2015.

[15] M. Schoeler and F. Worgotter, "Bootstrapping the Semantics of Tools: Affordance Analysis of Real World Objects on a Per-part Basis," *IEEE Trans. Cogn. Dev. Syst.*, vol. 8, no. 2, pp. 84–98, 2016.

[16] L. Riazuelo *et al.*, "RoboEarth Semantic Mapping: A Cloud Enabled Knowledge-Based Approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 2, pp. 432–443, 2015.

[17] H. Y. Gu, H. T. Li, L. Yan, and X. J. Lu, "A framework for Geographic Object-Based Image Analysis (GEOBIA) based on geographic ontology," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 40, no. 7W4, pp. 27–33, 2015.

[18] D. Arvor, L. Durieux, S. Andrés, and M. A. Laporte, "Advances in Geographic Object-Based Image Analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective," *ISPRS J. Photogramm. Remote Sens.*, vol. 82, pp. 125–137, 2013.

[19] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan, "An ontology-based retrieval system using semantic indexing," *Inf. Syst.*, vol. 37, no. 4, pp. 294–305, 2011.

[20] M: Wróbel, J. T. Starczewski, and C. Napoli, "Handwriting recognition with extraction of letter fragments", *International Conference on Artificial Intelligence and Soft Computing*, pp. 183-192, 2017.

[21] J. T. Starczewski, S. Pabiasz, N. Vladymyrska, A. Marvuglia, C. Napoli, and M. Woźniak, "Self organizing maps for 3D face understanding". *International Conference on Artificial Intelligence and Soft Computing*, pp. 210-217, 2017.