Chapter 22
Multilingual Image Corpus

Svetla Koeva

Abstract The ELG pilot project Multilingual Image Corpus (MIC 21) provides a large image dataset with annotated objects and multilingual descriptions in 25 languages. Our main contributions are: the provision of a large collection of high-quality, copyright-free images; the formulation of an ontology of visual objects based on WordNet noun hierarchies; precise manual correction of automatic image segmentation and annotation of object classes; and association of objects and images with extended multilingual descriptions. The dataset is designed for image classification, object detection and semantic segmentation. It can be also used for multilingual image caption generation, image-to-text alignment and automatic question answering for images and videos.

1 Overview and Objectives of the Pilot Project

Significant progress has been achieved in many multimodal tasks, such as image caption generation, aligning sentences with images in various types of multimodal documents and visual question answering. The shift of traditional vision methods challenged by multimodal big data motivates the creation of a new image dataset, the Multilingual Image Corpus (MIC21).

The MIC21 dataset is characterised by carefully selected images from thematically related domains and precise manual annotation for segmentation and classification of objects in over 20,000 images. The annotation is performed by drawing of or correcting automatically generated polygons, from which bounding boxes are automatically constructed. This allows for wide application of the dataset in various computer vision tasks: image classification, recognition and classification of single objects in an image or of all object instances in an image (semantic segmentation).

The annotation classes which are used belong to a specially designed ontology of visual objects which provides options for extracting relationships between objects in images; the construction of diverse datasets with different levels of granularity of

Svetla Koeva
Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria, svetla@dcl.bas.bg

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_22

object classes; and the compilation of appropriate sets of images illustrating different thematic domains. The ontology classes and their definitions, accompanied by illustrative examples, have been translated into 25 languages, which can be used for automatic interpretation of an image, caption generation and alignment of images with short texts such as questions and answers about the image content.

2 Methodology

We have divided the annotation process into four main stages: 1. definition of an ontology of visual objects; 2. collection of appropriate images; 3. automatic object segmentation and classification; and manual correction of object segmentation and manual classification of objects. The dataset contains four thematic domains (sport, transport, arts, security), which group highly related dominant classes such as *Tennis player*, *Soccer player*, *Limousine*, *Taxi*, *Singer*, *Violinist*, *Fire engine*, and *Police boat* in 130 subsets of images. We have used the COCO Annotator (Brooks 2019), which allows for collaborative work within a project, and offers tracking object instances and labelling objects with disconnected visible parts.

2.1 Ontology of Visual Objects

In current practice, WordNet is typically used in generating text queries for the creation of search-based image collections. For example, ImageNet uses 21841 synsets for image collection and their labeling (Russakovsky et al. 2015). A Visual Concept Ontology is proposed which organises concepts (Botorek et al. 2014), containing 14 top-level ontology classes divided into 90 more specific classes. Other datasets use a hierarchical organisation of object classes and mutually exclusive classes (Caesar et al. 2018), however, the number of concepts is usually relatively small.

The ontology of visual objects created for MIC21 embraces concepts that are thematically related and can be depicted in images. The four thematic domains (sport, transport, arts, security) are represented by 137 dominant classes, which show the main "players" within these domains. The ontology also embraces the hypernyms of the dominant classes up to the highest hypernym, which denotes a concrete object, and non-hierarchically related classes (called attributes) (Koeva 2021). The type of dominant class and the type of attribute class determine the type of the relation between them: *has instrument*, *wears*, *uses*, *has part*, etc. For example, the attribute classes for *Billiard player* are *Pool table*, *Billiard ball*, and *Cue*, while for *Bowler* – *Bowling alley*, *Bowl*, *Bowling pin*, *Bowling shoe* etc.; the hypernym classes for *Billiard player* and *Bowler* are *Player*, *Contestant* and *Person*.

Some of the classes and relations are inherited from WordNet (Miller et al. 1990). Additional classes and relations are included in the ontology in case they are not present in WordNet, for example *Bowler wears Bowling shoes*. Using the ontology of visual objects ensures the selection of mutually exclusive classes; the interconnectivity of classes by means of formal relations and an easy extension of the ontology with more concepts corresponding to visual objects.

2.2 Collection of Images and Metadata

The images in the dataset are collected from a range of repositories offering APIs: Wikimedia (images with Public Domain License or Non-copyright restrictions license)[1]; Pexels (images with a free Pexels license allowing free use and modifications)[2]; Flickr (images with Creative Commons Attribution License, Creative Commons Attribution ShareAlike License, no known copyright restrictions, Public Domain Dedication, Public Domain Mark)[3]; Pixabay (images with a free Pixabay license allowing free use, modifications and redistribution)[4]. The Creative Commons Search API is also used for searches on content available under Creative Commons licenses[5]. Over 750,000 images were collected in total and automatically filtered further by image dimensions, license types and for duplication. Each image is equipped with metadata description in JSON format: *filepath*; *source* (name of the repository or service used to obtain the image); *sourceURL* (URL of this repository or service); *license*; *author* (if available); *authorURL* (if available); *domain* (the domain the image belongs to); *width and height* (in pixels) etc.

3 Criteria for the Selection of Images

After the collection of images, we performed additional manual selection to ensure the quality of the dataset, applying the following criteria: i) The image has to contain a clearly presented object described by a given dominant class; ii) The object should (preferably) have no occluded parts; iii) The target object should be in its usual environment and in a position or use that is normal for its activity or purpose; iv) The instances of the target object in different images should not represent one and the same person, animal or artefact; v) Images with small objects, unfocused objects in the background or images with low quality are not selected; vi) Images which represent collages of photos or are post-processed are not selected.

The final selection of images is triple-checked independently by different experts: after the automatic collection, after the automatic generation of segmentation masks and during manual annotation.

[1] https://commons.wikimedia.org/wiki/Commons:Licensing

[2] https://www.pexels.com/license/

[3] https://www.flickr.com/services/developer/api/

[4] https://pixabay.com/service/license/

[5] https://api.creativecommons.org/docs/

3.1 Generation and Evaluation of Suggestions

To accelerate the manual annotation, an image processing pipeline for object detection and segmentation was developed. Two software packages – YOLACT (Bolya et al. 2019) and DETECTRON2 (Wu et al. 2019), and Fast R-CNN (Girshick 2015) models trained on the COCO dataset (Lin et al. 2014) were used for the generation of annotation suggestions. We also performed automatic relabelling for some of the predicted classes (usually for the dominant class and for some of its attribute classes), e. g., the COCO category *Person* within the subset *Golf* from the thematic domain *Sport* is replaced with the class *Golf player*. The performance of the models was evaluated over all domain-specific datasets within the domain *Sport* (see Figure 1).

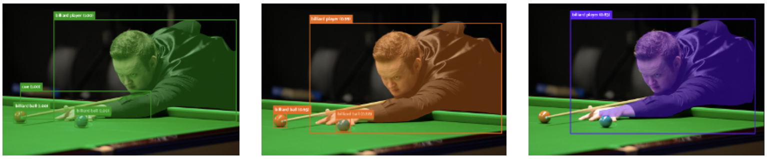

Fig. 1 Annotation results: human (left), YOLACT (middle) DETECTRON2 (right)

The results demonstrate similar behaviour with a slight predominance of one of the models, which was further used to predict the object classes in the datasets from the other three thematic domains. Altogether 253,980 segmentation masks were automatically generated, 194,212 of which were manually adjusted.

3.2 Annotation Protocol

The task for annotators was to outline polygons for individual objects in the image (either by approving or correcting the automatic segmentation or by creating new polygons) and to classify the objects against the classes from the predefined ontology. The annotation follows several conventions:

- An object within an image is annotated if it represents an instance of a concept included in the ontology.
- All objects from the selected dominant class and its attribute classes are annotated (for example, *Gondola* and the related objects *Gondolier* and *Oar*).
- If the object can be associated with different classes, this is recorded within the metadata (for example, for a female soldier – *Soldier* and *Woman*).

Quality control is provided by a second annotator who validates the implementation of the conventions and discusses the quality with the annotation group on a regular basis. If necessary, some of the images are re-annotated.

4 Multilingual Classes

For the purpose of the multilingual description of the images, all ontology classes have been translated into 25 languages: English (Princeton WordNet), Albanian, Bulgarian, Basque, Catalan, Croatian, Danish, Dutch, Galician, German, Greek, Finnish, French, Icelandic, Italian, Lithuanian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Spanish, and Swedish.

Openly available wordnets have been used from the Extended Open Multilingual WordNet.[6] For the ontology classes which are not inherited from WordNet the appropriate WordNet hypernyms are used. Where WordNet translations are not available, additional sources of translations as BabelNet[7] are employed. The multilingual translations of classes are presented in a separate JSON file which contains information about the language and the translation source. The translations of the ontology classes are accompanied by their synonyms, the concept definition and usage examples (if available in the sources).

5 Conclusions and Results of the Pilot Project

The Multilingual Image Corpus provides fully annotated objects within images with segmentation masks, classified according to an ontology of visual objects, thus offering data to train models specialised in object detection, segmentation and classification (Table 1). The ontology of visual objects allows easy integration of annotated images in different datasets as well as learning the associations between objects in images. The ontology classes are translated into 25 languages and supplied with definitions and usage examples. The explicit association of objects and images with appropriate text fragments is relevant for multilingual image caption generation, image-to-text alignment and automatic question answering for images and video.

Domain	Subsets	Number of Images	Number of Annotations
Sport	40	6,915	65,482
Transport	50	7,710	78,172
Arts	25	3,854	24,217
Security	15	2,837	35,916
MIC21	**130**	**21,316**	**203,797**

Table 1 Multilingual Image Corpus: basic statistics

[6] http://compling.hss.ntu.edu.sg/omw/summx.html

[7] https://babelnet.org/guide

All annotations and image metadata are available for commercial and non-commercial purposes in accordance with the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Bolya, Daniel, Chong Zhou, Fanyi Xiao, and Yong Jae Lee (2019). "YOLACT: Real-time Instance Segmentation". In: pp. 9156–9165. DOI: 10.1109/ICCV.2019.00925.

Botorek, Jan, Petra Budíková, and Pavel Zezula (2014). "Visual Concept Ontology for Image Annotations". In: *CoRR*. URL: http://arxiv.org/abs/1412.6082.

Brooks, Justin (2019). *COCO Annotator*. URL: https://github.com/jsbroks/coco-annotator/.

Caesar, Holger, Jasper Uijlings, and Vittorio Ferrari (2018). "COCO-Stuff: Thing and Stuff Classes in Context". In: *Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218.

Girshick, Ross (2015). "Fast R-CNN". In: pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.

Koeva, Svetla (2021). "Multilingual Image Corpus: Annotation Protocol". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. INCOMA, pp. 701–707.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár (2014). "Microsoft COCO: Common Objects in Context". In: *European Conference on Computer Vision (ECCV)*. Zürich, pp. 740–755.

Miller, George, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (1990). "Introduction to WordNet: An on-line lexical database". In: *International Journal of Lexicography* 3, pp. 235–244.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 116, pp. 157–173.

Wu, Yuxin, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick (2019). *Detectron2*. URL: https://github.com/facebookresearch/detectron2.