Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1509–1518
Marseille, 20-25 June 2022

Multilingual Image Corpus – Towards a Multimodal and Multilingual Dataset

Svetla Koeva[1], **Ivelina Stoyanova**[2], **Jordan Kralev**[3]
[1,2,3] Institute for Bulgarian Language, Bulgarian Academy of Sciences
[3]Technical University, Sofia
[1]svetla@dcl.bas.bg, [2]iva@dcl.bas.bg, [3]kralev@dcl.bas.bg

Abstract

One of the processing tasks for large multimodal data streams is the automatic image description (image classification, object segmentation and classification). Although the number and the diversity of image datasets is constantly expanding, there is still a huge demand for more datasets in terms of variety of domains and object classes covered. The goal of the project Multilingual Image Corpus (**MIC21**) is to provide a large image dataset with annotated objects and object descriptions in 25 languages. The Multilingual Image Corpus relies on an Ontology of Visual Objects (based on WordNet) and comprises a collection of thematically related images whose objects are annotated with segmentation masks and labels linked to the ontology classes. The dataset is designed both for image classification and object detection and for semantic segmentation. The main contributions of our work are: a) the provision of a large collection of high-quality images licensed for commercial and non-commercial use; b) the compilation of the Ontology of Visual Objects based on WordNet noun hierarchies; c) the automatic object segmentation within the images followed by precise manual editing and the annotation of object classes; and d) the mapping of objects and images to extended multilingual descriptions based on WordNet inner- and interlingual relations. The dataset can be used also for multilingual image caption generation, image-to-text alignment and automatic question answering for multimedia content.

Keywords: multilingual image corpus, multilingual dataset, multimodal dataset

1. Introduction

Nowadays the multimodal content combining text, images and video is rapidly increasing. The processing of big multimodal data based on reliable computer vision models poses ever increasing challenges on the quality and the amount of the training data. Although the size of the image datasets can be artificially increased by synthetic data, the synthetic data has to be grounded by real-world images and their annotations.

The multimodal data sources supply richer information than a single-mode input. Advanced technological solutions for analysing multimodal content can be provided using computer vision methods together with natural language processing. There has been significant progress in many multimodal tasks such as image caption generation (Xu et al., 2016), aligning sentences with images in various types of multimodal documents (Hessel et al., 2019) and visual question answering (Antol et al., 2015; Sethuraman et al., 2021; Le et al., 2020; Gao et al., 2020). These tasks can be achieved by interpreting an image and the corresponding short text such as a caption, question or description of the objects in the image.

The shift of traditional data fusion methods challenged by big multimodal data motivates the creation of a new image corpus, the Multilingual Image Corpus (**MIC21**)[1], which is characterised by carefully selected images that illustrate a set of related domains, as well as by precise manual annotation for segmentation and classification of objects in the images. The dataset is designed both for image classification and object detection and for semantic segmentation.

MIC21 has the following main features: (a) it is a large dataset containing thousands of images and annotations; b) the annotation classes belong to a specially designed Ontology of Visual Objects; c) the object segmentation and multi-label classification according to the Ontology classes is evaluated by experts; d) the multilingual description of the objects in the images includes names of object classes and attached definitions of concepts presented in 25 languages.

The Ontology of Visual Objects provides options for extracting: relationships between annotated objects; diverse datasets with different levels of granularity of object classes; appropriate sets of images illustrating different thematic domains. Last but not least, the use of the Ontology of Visual Objects allows the expansion of the dataset depending on the specific requirements of scientific or commercial projects.

The annotation is performed by drawing or adjusting automatically generated polygons, from which bounding boxes are also automatically constructed. This allows for a wide range of application of the dataset in the field of computer vision: image classification, recognition and classification of single objects within an image or of all instances of an object in an image, the so-called semantic segmentation.

The dataset can be used also for multilingual image caption generation, as the labels of the Ontology classes are presented in 25 languages. The multilingual description is based on wordnets for the respective lan-

[1] https://dcl.bas.bg/mic21/

guages, further expanded from other sources since the wordnets have different scope and coverage. Applications of the dataset also extend to the fields of alignment between images and sentences within a text and automatic question answering for images and videos. For this purpose, both the multilingual definitions of the concepts and the inferences that can be made on the basis of the relations and axioms defined within the Ontology of Visual Objects can be exploited.

The paper is organised as follows. Section 2 presents a brief comparison between our work and image datasets whose compilation or annotation was based on WordNet. Section 3 describes the collection of images, followed by the conventions for image selection (section 4) and their metadata description (section 5). Section 6 is dedicated to the Ontology of Visual Objects with its classes, relations and axioms. Section 7 outlines the automatic prediction of objects and their classes which facilitates the manual annotation. The automatic prediction is then compared to the manually edited annotation (section 8) and the annotation conventions are discussed (section 9). The multilingual description of the Ontology classes (section 10) is followed by the terms and conditions of the copyright of the dataset (section 11).

2. Related Work

There are several datasets which have been widely used as a benchmark for object detection, semantic segmentation and classification tasks. Only a few of them use ontologies or ontology-like resources for object classification.

VQA[2] is a dataset containing open-ended questions about images. These questions require computer vision capabilities, language and commonsense knowledge to answer. 265,016 images (COCO and abstract scenes) are mapped with at least 3 questions (5.4 questions on average) per image, 10 ground-truth answers per question and 3 plausible (but likely incorrect) answers per question (Goyal et al., 2017).

LabelMe[3] is a dynamically developing dataset which contains hundreds of thousands of polygon annotations, thousands of static images and sequence frames with at least one labelled object (Russell et al., 2008). A particular feature of this collection is that it is being developed by users who can add images and categories and can annotate uploaded images. This option however may result in some level of inconsistency based on the decisions of the different users about the annotation protocol. The WordNet noun synonym sets (synsets) are used to extend the categories, to avoid inconsistency by means of manual editing and to unify the descriptions provided by different users.

ImageNet[4] is still one of the largest datasets and it has been used for the development of many computer vision applications. ImageNet uses WordNet noun hierarchies for image collection and labelling. 21,841 WordNet synsets within the WordNet hierarchy are depicted in 14,197,122 images (as of August 2014) (Russakovsky et al., 2015). The dataset has been used in multiple competitions on tasks such as: image classification (algorithms produce a list of object categories present in the image), single-object localisation (algorithms produce a list of object categories present in the image, along with an axis-aligned bounding box indicating the position and the scale of one instance of each object category), object detection (algorithms produce a list of object categories present in the image along with an axis-aligned bounding box indicating the position and the scale of every instance of each object category) (Russakovsky et al., 2015).

The taxonomic organisation of nouns in **WordNet** allows for different level of granularity (using more abstract or fine-grained categories) when describing objects. WordNet is a semantic network comprising nodes (each node hosts synonyms denoting the same concept) and arcs connecting the nodes and encoding different types of relations (semantic: genus-species, part-whole, etc.; extralinguistic: membership in a thematic domain; inter-language: translation equivalents) (Miller et al., 1990). However, the hierarchy of nouns in WordNet is based on the organisation of concepts in the human mind and does not correspond exactly to the Ontology of Visible Objects (section 6).

The **COCO (Microsoft Common Objects in Context)** dataset[5] (Lin et al., 2014) contains more than 328,000 images with manually annotated instances of objects (2.5 million). The dataset has had several releases since 2014 and provides data for object detection, segmentation, keypoint detection and captioning. The different parts of the dataset are annotated with bounding boxes (for object detection) and instance segmentation masks with 80 object categories; natural language descriptions of the images; keypoints (17 possible keypoints such as *left eye, nose*); per-pixel segmentation masks with 91 stuff categories such as *grass, wall*; full scene segmentation with 80 thing categories such as *person, bicycle, elephant*; dense pose – for each labelled person image pixels are mapped to a template 3D model.

In current practice, WordNet is usually used in generating text queries for compilation of search based image collections. A **Visual Concept Ontology** is proposed which organises visual concepts (abstract notions or objects that are typically depicted in photos) (Botorek et al., 2014). For the construction of the Visual Concept Ontology over 400 "significant" noun synsets (that have at least 300 hyponyms) are extracted from WordNet, then synsets with very "general" meaning such as *entity* or *thing* were removed. This results in 14 top-level ontology classes, which are divided further into 90 more specific classes. Semantically similar synsets

[2] https://visualqa.org/index.html

[3] http://labelme.csail.mit.edu/Release3.0/

[4] https://www.image-net.org

[5] https://cocodataset.org

are merged into a common class and additional links are established between semantically related synsets such as *roof* and *house*. The Ontology excludes some abstract concepts and the proper names that are part of WordNet, however it is limited to a small number of concepts that correspond to visual objects.

In the largest collection of datasets available on the internet (1699 image datasets are listed as of April 2022)[6] the datasets are provided with descriptions and links to the sources and related papers. Among them 137 datasets are designed for semantic segmentation; 128 – for image classification; 131 – for object detection; 18 datasets provide polygon annotations.

To summarise, the tendency in image annotation is to shift from small training datasets to large-scale collections which require crowdsourcing in order to engage large volume of human effort. Although the number and the diversity of image datasets is constantly expanding, there is still a huge demand for more datasets in terms of variety of domains and object classes covered. **MIC21** is one of the few datasets in which a large number of classes are used to describe objects. Moreover, the classes are organised in an Ontology of Visual Objects and thematically related objects are linked within the Ontology, which facilitates the drawing of conclusions about the scenes depicted in the images and the solving of various tasks related to text and image processing. The presentation of classes, their definitions and usage examples in 25 languages provides opportunities for multilingual and multimodal processing.

3. Collection of Images

The images in the dataset are collected automatically from a range of repositories:

(1) **Wikimedia** is a collection of 78 million media files – images, videos, etc. Images are distributed under the Public Domain License or Non-copyright restrictions license.[7]

(2) **Pexels** is a large repository of images, offering a free API for automatic collection using search queries. All images are distributed with a free Pexels license allowing free use, modifications and commercial use.[8]

(3) **Flickr** is a large repository for media content with over 5 billion photos. Flickr offers an API for automatic image searches with a limited number of queries per hour (3,600). The images are distributed with a variety of licenses. We use only a selected number of non-restrictive licenses: Creative Commons Attribution License, Creative Commons Attribution ShareAlike License, No known copyright restrictions, Public Domain Dedication (CC0), Public Domain Mark.[9]

(4) **Pixabay** is a large repository of images, also offering a free API for automatic collection using search queries. All images are distributed with a free Pixabay license allowing free use, modifications and redistribution, including for commercial use. [10]

(5) **Creative Commons Search API** is a service allowing searches on content available under the Creative Commons licenses. It offers an opportunity for more precise searches over a large number of sources, including some of the above. Images sourced using the CC Search API are labeled by both the destination repository (they are found in) and the CC API as the method of collection.[11]

(6) A small number of images were collected semi-automatically (manual queries and automatic downloading of images) from smaller repositories allowing free use with attribution, redistribution and commercial use: **Pxfuel**,[12] **Public Domain Pictures**.[13]

Each repository was accessed through an API and search queries were designed to search for images using a single term or a phrase. The queries targeted the dominant classes formulated in the Ontology of Visual Objects. The most common challenges for collecting images for the dataset are:

• Using general query terms (e.g. *cricket*) returned a large number of low relevant results which required time-consuming manual selection.

• Using very specific queries (e.g., *cricket fine leg*) returns a small number of highly relevant but insufficient results.

In order to balance these out, for each category we define a series of queries for specific terms and combinations of a general and a specific term. All collected results are then filtered by image dimensions and license type (images of low quality and/or distributed under restrictive licenses are discarded). Some images appeared in more than one repository which necessitated additional checks for duplication removal, as well as double-checking of copyright terms. We have collected over 750,000 images in total.

4. Criteria for Selection

After the automatic collection of images, we perform additional manual selection to ensure the quality of each domain-specific dataset. The following criteria for selection are observed (Koeva, 2021):

• The image has to contain a clearly presented object described by a given dominant class.

• (Preferably) the object should not have occluded parts. If parts of the object are occluded, they should not be essential for its recognition.

• The target object should be in its usual environment, in a usual position and in use corresponding to its purpose.

• The target object should be represented with its inherent attributes.

[6] https://paperswithcode.com/datasets

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

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

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

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

[11] https://api.creativecommons.engineering/v1/#operation/image_search

[12] https://www.pxfuel.com/

[13] https://www.publicdomainpictures.net/

• The target object should be in different positions, photographed from diverse viewpoints and angles and the object background should vary to a sufficient degree.
• The instances of the target object should not represent one and the same person, animal or artefact.
• (Preferably) images with up to 10 objects are selected (the objects can belong to different classes or can be instances of one and the same class). If there are images with only one object, then it should belong to a dominant class.
• Images with small objects, unfocused objects in the background or images of low quality (low resolution, blurriness caused by an out-of-focus lens, low illumination level, etc.) are not selected.
• Images which represent collages of photos, drawings 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 after the manual validation and editing of segmentation masks, new polygon outlines and assignment of appropriate classes. Table 1 shows the final selection of images distributed by source.

Source	Number of images
Wikimedia	10,312
Pexels	3,838
CC Search API	3,092
Pixabay	2,686
Flickr	1,365
Other	23

Table 1: Distribution of images by source

5. Metadata

MIC21 contains subdatasets annotated for the following thematic domains: **Sport**, **Transport**, **Arts**, and **Security**, where each domain is further represented by subdomains such as **Tennis player**, **Limousine**, **Photographer** and **Policeman**.
Each image is supplied with metadata in JSON format containing the following components:
• Image file data: **file path** (within the dataset), **file name**, **image id**;
• Subdataset data: **subdataset name**, **subdataset id**;
• Image size: **original width / height** (in pixels);
• Image source data: **source name**, **source url**, **source url last access**, **source tags** (whenever available, keywords in the image description obtained from the original source and/or words used for the query which produced the image), **author name or username**, **author url** (if available);
• Image licensing data: **image license** (type of license applied to the original image);
• Main dataset information: **MIC21 license**, **MIC21 provider**, **MIC21 provider url**, **MIC21 provider contact**, **MIC21 project url**, **MIC21 credit**.

6. Ontology of Visual Objects

It was pointed out that different knowledge representations share the following minimal set of components (Corcho et al., 2006): **concepts**, which represent sets or classes of entities in a thematic domain; **relations** between concepts; **instances**, which represent the actual entities (individuals); and **axioms**, which represent facts that are always true in the topic area of the ontology.
We adopted the following definition (Bozsak et al., 2002): An ontology is a structure

$$O := (C, \leq_C, R, \leq_R)$$

consisting of (i) two disjoint sets C and R called concept identifiers and relation identifiers respectively, (ii) a partial order $\leq_C$ on C called concept hierarchy or taxonomy, (iii) a function $\sigma : R \to C \times C$ called signature and (iv) a partial order $\leq_R$ on R called relation hierarchy.
The proposed Ontology of Visual Objects includes concepts which are characteristic for the thematic domains of **Sport**, **Transport**, **Arts**, and **Security**. **MIC21** contains a total of 130 smaller datasets pertaining to different subdomains, each of which can be classified to one of the four main ones, for example, **Chess** and **Pole vaulting** are subdomains of **Sport**, while **Sedan** and **Double-decker** – of **Transport**, and so on. The choice of thematic domains and subdomains is motivated by two main factors:
(1) The images should contain objects which could be automatically recognised and labeled with upper-level classes (for example, *man* and *car*) which then could be sub-classified as *chess player*, *pole vaulter*, *sedan* and *taxi*;
(2) There should be a sufficient number of appropriate images available to illustrate objects from the selected thematic subdomains.
The Ontology of Visual Objects contains classes, relations and axioms.

6.1. Classes

More than half of the **classes** correspond to WordNet concepts which can be represented by visual objects (485 out of 851 classes). Among the classes we made a differentiation between dominant classes and attribute (contextual) classes.
Each thematic domain is represented by several **dominant classes** which show some of the main "players" within this domain differentiated by their type or their function. For example, the dominant classes for the domain **Security** are: *policeman, soldier, fireman*, etc. For the definition of the dominant classes, we use the WordNet sister hyponyms at a certain level (the lowest level allowing classification without specific knowledge for the domain). The selected dominant classes

for all thematic domains in focus are 137, with related 145 parent and 569 attribute classes.

For each dominant class a parent class is selected from the WordNet noun hierarchies and parent classes are extracted recursively up to the final class which represents a visual object. For example, classes like *basketball player, acrobat, football player*, etc. are hyponyms of *athlete* 'a person trained to compete in sports'. *Athlete* in its turn is a hyponym of *contestant* 'a person who participates in competitions', which is a hyponym of *person*. However, the hypernym of *person* is *organism*, an abstract notion, which is not included in the Ontology. As a result of this approach, thousands of annotations can be assigned to objects representing small number of classes, while the annotations with more general classes will be inherited automatically.

The **attributes** in the ontology are classes related to the dominant ones. The type of the dominant class and the type of the attribute class determine the type of the relation between them which expresses a specific property attribution: **has instrument, wears, uses, has part**, etc. For example, the attribute classes for *cricketer* are *cricket bat, cricket ball, cricket helmet, wicket* and *referee*, while for *climber – climbing helmet, chalk bag, climbing backpack*, and so on.

Each class in the Ontology is represented by a unique label which in most cases is one of the synonyms in the corresponding WordNet synset (in case of ambiguity, a descriptive label is constructed).

6.2. Relations

The **relations** used in the Ontology are relations between classes. Part of the relations and their properties are inherited from WordNet (**is a** and **has part**).

Relations between dominant and attribute classes are not hierarchical. For the linking of attribute classes, we use one WordNet relation – **has part** and 13 relations that are not overtly represented in WordNet and are additionally created for the Ontology (for example, **wears**, **is next to** and **plays with**).

Altogether, 15 relations are used in the Ontology, with 828 instances of the **is a** relation; 241 instances of the **wears** relation, 210 instances of the **has part** relation, and so on.

6.3. Axioms

Axioms serve to model sentences that are always true (Gruber, 1995) and they can be used to infer new knowledge. An axiom system for an ontology is a pair (AI, α) where (i) AI is a set whose elements are called axiom identifiers, and (ii) α is a mapping. The elements of $A := \alpha(AI)$ are called axioms (Cimiano and Handschuh, 2003).

Using the Ontology of Visual Objects ensures the selection of mutually disjoint classes as annotation labels, built-in interconnectivity of classes by means of formal relations and an easy extension of the proposed ontology with more concepts corresponding to visual objects.

The images are collected, annotated and supplied with multilingual descriptions according to the Ontology classes into which image objects can be classified.

7. Generation of Annotation Proposals

The manual annotation tasks are performed using the COCO annotator[14] which allows for simultaneous work on a project and offers useful functionalities that facilitate image annotation: tracking object instances, labelling objects with disconnected visible parts, etc.

To accelerate manual annotation, we have developed an image processing pipeline for object detection and object segmentation using pre-trained models developed for the original COCO labeling domain.[15] Two software packages are used in the processing pipeline – YOLACT (Bolya et al., 2019) with *Resnet50-FPN* backbone and Detectron2 (Wu et al., 2019) using Fast-RCNN architecture with Resnet backbone (He et al., 2015).

Recent developments in computer vision allow near real-time object localisation and classification within video or images through the application of multi-layer network architectures – R-CNN (Girshick et al., 2014), YOLO (Bolya et al., 2019), SPPNet (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), etc. The general structure of the image processing framework based on the Fast R-CNN is presented on Figure 1.

Figure 1: Fast R-CNN structure

The image dataset is stacked in an input data batch where each element of the batch contains image pixel data as a three-dimensional array (tensor), followed by image attributes (width and height) and ground-truth training data, if available. The ground-truth data for image objects includes a bounding box, an object mask and a class label.

A large number of ready-to-use Fast R-CNN models are trained on the COCO dataset, where the objects in the input images are classified into 80 COCO categories. The result from the initial dataset processing is the instance segmentation and classification into the COCO domain categories stored in MS COCO JSON format.

After the automatic prediction of object classes we performed relabelling of the initially assigned COCO categories with the appropriate **MIC21** classes. For example, the COCO category *person* within the domain-specific dataset **Gymnastics** is replaced with the **MIC21** class *gymnast*, while in the domain-specific dataset **Dancer** – with the **MIC21** class *dancer*.

The manual work of human annotators is considerably reduced to adjusting the boundaries of automatically generated annotation masks and editing class labels.

[14] https://github.com/jsbroks/coco-annotator

[15] https://github.com/DCL-IBL

For the Ontology classes with no corresponding COCO categories the annotators have to create new polygons and to classify objects. Some incorrectly generated segmentation masks have to be deleted. Altogether, 253,980 segmentation masks were automatically generated, from which 194,212 were subsequently manually adjusted or deleted.

The image processing pipeline also includes a toolset for post-processing operations on the domain-specific datasets such as merging or splitting datasets and removing or replacing class labels.

8. Evaluation of Annotation Proposals

The evaluation of the performance of some pre-trained models for predicting and labeling objects in images was necessary in order to decide which of the models is most effective for our purposes. For the evaluation we used the ground-truth annotations produced by human experts for images in the thematic domain **Sport**. The frameworks which we used – YOLACT and Detectron2, offer selections of pre-trained models over the COCO dataset for object detection and image segmentation. Converting modules were developed to interface the models inputs and outputs to the presentation layer (Brooks, 2019; Moore and Corso, 2020) in the required JSON format. In order to carry out evaluation, we performed static one-to-one mapping (where appropriate) between the COCO categories and the domain-specific **MIC21** classes. For example, the COCO category *person* is common for all subdatasets and is mapped to domain specific classes such as *billiard player*, *soldier*, etc.

We use the standard evaluation metrics for object classification projects – precision, recall and F_1-score. In order to calculate them for each image, we determine the number of:

• true positives (correct labeling) – N_{tp},

• true negatives (correctly dropped irrelevant objects) – N_{tn},

• false positives (falsely detected irrelevant objects) – N_{fp},

• and false negatives (undetected relevant objects) – N_{fn}.

The precision metric P shows what fraction of the correctly detected objects in the image are correctly labelled (or classified):

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}}, \tag{1}$$

and the recall metric R shows the fraction of correctly detected objects from all ground-truth objects:

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}}. \tag{2}$$

Both metrics are equally important, so we use the harmonic mean $2PR/(P + R)$. For an illustration, the resultant average precision, recall and F_1-scores for the **Billiard** subdataset processed with an YOLACT model are presented in Table 2.

Label	Prec.	Recall	$\mathbf{F_1}$	Support
billiard ball	0.86	0.46	0.60	989
billiard player	0.27	0.97	0.42	232
cue	0.00	0.00	0.00	246
man	0.00	0.00	0.00	14
pool table	0.00	0.00	0.00	231
woman	0.00	0.00	0.00	11

Table 2: The average metric for the **Billiard** subdataset using the YOLACT model

The COCO categories *ball* and *person* are directly mapped to domain-specific classes *billiard ball* and *billiard player* respectively, giving rise to non-zero precision and recall values. On the other hand, other domain-specific classes such as *cue*, *man*, *pool table* and *woman* are not present in the COCO ontology, hence, they participate with 0 true positives.

The performance of the YOLACT and Detectron2 models can be evaluated for all classes for which a correspondence with the COCO categories could be established. For example, the class *billiard ball* is correctly recognised in 86% of the cases and 46% of the ground-truth instances labeled for this class are caught by the model. For the class *billiard player* 97% of ground-truth instances are correctly covered by the model, however, with a lot of false positives leading to precision value of 27%. The last column from Table 2 shows the number of labeled instances in the ground-truth dataset.

The performance of the two models was evaluated over all domain-specific datasets within the domain **Sport** (an example is demonstrated on Figure 2). The analysis of models' performance motivated their selection for further prediction of object classes in the datasets.

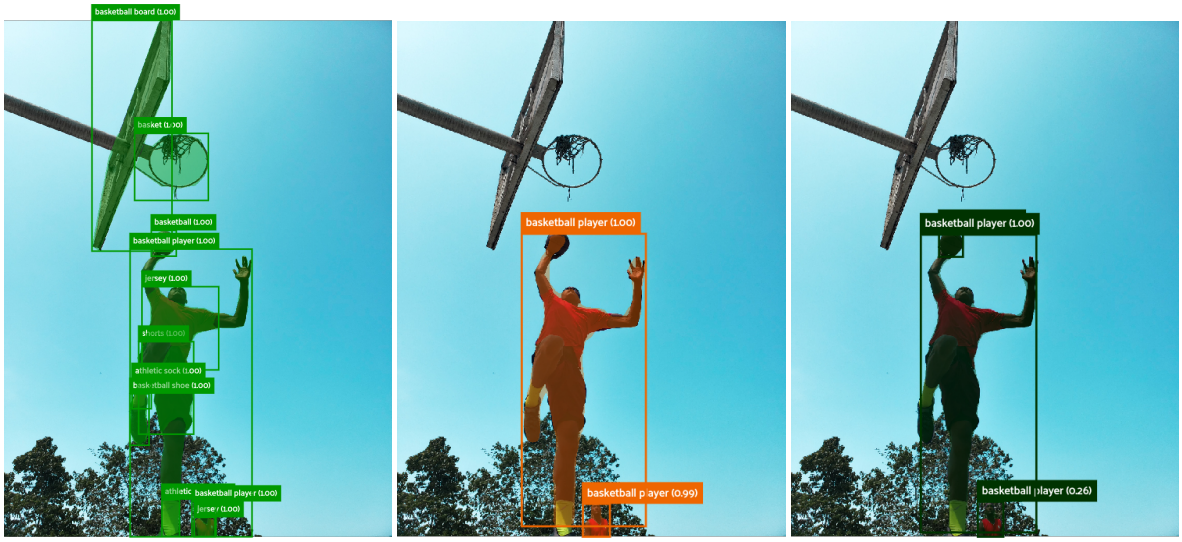
9. Annotation Protocol

The manual annotation consists of outlining polygons for individual objects in the image (either by confirming or editing the automatic segmentation or by creating new polygons) and classifying the objects into the classes from the predefined Ontology. The annotation follows several conventions:

• An object displayed 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 attribute classes related with it are annotated (for example, the class *tennis player* and the related objects *racket* and *tennis ball*; *chess player* and the related objects *chessman*, *chess board*, and *clock*).

• If the object can be additionally associated with a different class, this is recorded within the metadata field (for example, if the *climber* is not a *man* but a *boy*, a



(a) Human (b) Detectron2 (c) YOLACT

Figure 2: Example of ground-truth annotation and annotation results using Detectron2 and YOLACT

woman or a *girl*, the additional class is added to the metadata).

Quality control procedures are carried out by a second annotator who validates the annotations and ensures the implementation of the conventions. The quality issues are discussed within the annotators' team weekly. If necessary, some of the images are re-annotated.

10. Multilingual Classes

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

In providing translation equivalents to Ontology classes, priority is given to WordNet. There are wordnets for many languages which are all linked to the Princeton WordNet (Bond and Foster, 2013), and in this way are linked to each other. To this end we employ openly available wordnets from the Extended Open Multilingual Wordnet project[16] or official distribution webpages.

More than half of the Ontology class labels have been matched initially to synsets in the Princeton WordNet, and then to other languages. From the synsets in different languages we extracted the synonyms of Ontology classes, the definitions of the concepts and usage examples. For example, the concept *motorcycle* is mapped to the synset *motorcycle*:1; *bike*:1, and then to synsets in other languages (Table 3).

Where WordNet translations are not available, some additional sources of translations are employed:

Language	Synset
EN (Princeton)	eng-30-03790512-n *motorcycle*, *bike*
French	*vélo, motocyclette, motocycle, moto*
Spanish	*moto*, *motocicleta*, *motociclo*
Croatian	*motocikl*, *motor*, *motorkotač*
Italian	*motociclo*
Polish	*motocykl*, *motor*

Table 3: Multilingual representation of Ontology classes

• For classes that are linked to WordNet synsets but that were not present in the wordnet for a given language, its BabelNet translation was provided if available.[17] BabelNet is used to extract the definition of the class if it is not available in the respective wordnet.

• Further, for some of the cases where an Ontology class is not directly mapped to a WordNet synset, BabelNet is used for the extraction of translation equivalents and definitions.

• For the other languages, if translation equivalents could not be extracted from neither WordNet nor BabelNet, we apply machine translation with subsequent automatic ranking to select the most suitable translation candidate. The ranking is based on the frequency of a translation equivalent appearing when translating independently from English and from Bulgarian to the target language, as well as the frequency of obtaining the source label when translating back.

• For the cases where no suitable translation equivalent is found, we use the parent class from the Ontology (in $< 1\%$ of cases).

The multilingual translations of classes are presented

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

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

in a separate JSON file. Each translation includes the language, the label (dominant synonym considered to be the most frequent or descriptive and picked as a label in the Ontology), a list of synonyms in that language, translation source, definition and usage examples (if available). The coverage of Ontology classes across several languages and the sources they are extracted from are shown in Table 4. Definitions of concepts are best covered for English and Bulgarian (over 81.0%), with another 3 languages with over 75% and further 14 languages with over 50% of definitions. Usage examples appear for only just over a half of the languages with the best coverage in Bulgarian (38% of classes).

Language	% WN	% manual	% BN	% MT
English	57.0	43.0	–	–
Bulgarian	57.0	43.0	–	–
Romanian	44.6	–	17.3	38.1
Spanish	43.5	–	19.0	37.5
French	37.3	–	38.0	24.7
Italian	36.9	–	18.7	44.4

Table 4: Coverage of the Ontology classes in several languages: from WordNet (WN), manually defined, from BabelNet (BN), or using machine translation (MT)

In the last decade some multilingual multimodal datasets have been made available such as ImageNet (Russakovsky et al., 2015), BabelPic (Calabrese et al., 2020), MultiSubs (Wang et al., 2021), MMID image/word dataset (Hewitt et al., 2018). The main advantages of **MIC21** over well known large-scale multimodal and multilingual resources are:
(a) labelling of related and characteristic domain-specific objects, which facilitates domain-specific knowledge extraction from multimodal data;
(b) expandability – the knowledge is presented in multiple languages in a structured manner and the methodology can be applied for any number of domains and languages. A new language can be linked through translation from English or any other language in the dataset.

11. Copyright

All **MIC21** annotations and metadata are available for commercial and non-commercial purposes to be downloaded, copied, modified, distributed, displayed and used in accordance with the Creative Commons Attribution-ShareAlike 4.0 International License.[18]
The original images included in the dataset are distributed with the following licenses: Creative Commons licenses (Attribution, Attribution-ShareAlike, Attribution-NoDerivs, Public Domain Dedication (CC0), Public Domain Mark);[19] Pexels license[20] and Pixabay license[21] (via official APIs): royalty-free right to use, download, copy, modify, adapt and redistribute the Content for commercial or non-commercial purposes.
MIC21 is available for download at the ELG platform.[22]

12. Conclusions

We introduced the **Multilingual Image Corpus (MIC21)**, aimed at facilitating research on multilingual and multimodal data processing. **MIC21** provides pixel-level annotations for the selected dominant classes and their parent and attribute classes in four thematic domains, thus offering data to train models specialised in object detection, segmentation and classification in these domains. The current state of **MIC21** is presented in Table 5.

Domain	Subdomains	Images	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**	**203,787**

Table 5: The Multilingual Image Corpus in Numbers

The selected classes for annotation are organised in an Ontology of Visual Objects that affords options to compile different datasets with respect to a wide range of tasks. The ontological organisation of object classes provides data for learning associations between objects in images, for identifying relations between objects and for aligning objects and relations with text fragments. The labels of object classes are linked to their synonyms, definitions and usage examples in 25 languages. The multilingual layer makes the dataset suitable for artificial intelligence applications (e.g., multilingual image captioning, question-answering, machine translation of multimodal content).

13. Acknowledgements

The **Multilingual Image Corpus (MIC21)** project was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

[18] hhttps://creativecommons.org/licenses/by-sa/4.0/
[19] https://creativecommons.org/
[20] https://www.pexels.com/terms-of-service/
[21] https://pixabay.com/service/terms/
[22] https://live.european-language-grid.eu/catalogue/corpus/18029

14. Bibliographical References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C., and Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015)*, pages 2425–2433, Los Alamitos, CA, USA, dec. IEEE Computer Society.

Bolya, D., Zhou, C., Xiao, F., and Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pages 9157–9166.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Botorek, J., Budíková, P., and Zezula, P. (2014). Visual Concept Ontology for Image Annotations. *Computing Research Repository (CoRR), arXiv*. http://arxiv.org/abs/1412.6082.

Bozsak, E., Ehrig, M., Handschuh, S., Hotho, A., Maedche, A., Motik, B., Oberle, D., Schmitz, C., Staab, S., Stojanovic, L., et al. (2002). KAON – towards a large scale Semantic Web. In *International Conference on Electronic Commerce and Web Technologies*, pages 304–31. Springer.

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

Calabrese, A., Bevilacqua, M., and Navigli, R. (2020). Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts. In *Proceedings of ACL*, pages 4680–4686.

Cimiano, P. and Handschuh, S. (2003). Ontology-based Linguistic Annotation. In *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, pages 14–21, Sapporo, Japan, jul. Association for Computational Linguistics.

Corcho, Ó., Fernández-López, M., and Gómez-Pérez, A. (2006). Ontological engineering: Principles, methods, tools and languages. In Coral Calero, et al., editors, *Ontologies for Software Engineering and Software Technology*, pages 1–48. Springer.

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5):829–864, 05.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Computing Research Repository (CoRR), arXiv*. http://arxiv.org/abs/1311.2524.

Girshick, R. (2015). Fast R-CNN. *Computing Research Repository (CoRR), arXiv*. http://arxiv.org/abs/1504.08083.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2017)*.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43:907–928, December.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *Computing Research Repository (CoRR), arXiv*. http://arxiv.org/abs/1512.03385.

Hessel, J., Lee, L., and Mimno, D. (2019). Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. *Computing Research Repository (CoRR), arXiv*. http://arxiv.org/abs/1904.07826.

Hewitt, J., Ippolito, D., Callahan, B., Kriz, R., Wijaya, D. T., and Callison-Burch, C. (2018). Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Koeva, S. (2021). Multilingual Image Corpus: Annotation Protocol. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 701–707, Held Online, September. INCOMA Ltd.

Le, T. M., Le, V., Venkatesh, S., and Tran, T. (2020). Hierarchical Conditional Relation Networks for Video Question Answering. *Computing Research Repository (CoRR), arXiv*. https://arxiv.org/abs/2002.10698.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, Zürich.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Moore, B. E. and Corso, J. J. (2020). Fiftyone. *GitHub*. https://github.com/voxel51/fiftyone.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 116:157–173.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173.

Sethuraman, M. G., Payani, A., Fekri, F., and Kerce, J. C. (2021). Visual Question Answering based

on Formal Logic. *Computing Research Repository (CoRR), arXiv*. https://arxiv.org/abs/2111.04785.

Wang, J., Madhyastha, P., Figueiredo, J., Lala, C., and Specia, L. (2021). MultiSubs: A Large-scale Multimodal and Multilingual Dataset. *Computing Research Repository (CoRR), arXiv*. https://arxiv.org/abs/2103.01910.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. *GitHub*. https://github.com/facebookresearch/detectron2.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2016). Show, attend and tell: Neural image caption generation with visual attention. https://arxiv.org/abs/1502.03044.