CEUR-WS.org/Vol-2846/paper39.pdf

Towards More Robust Fashion Recognition by Combining Deep-Learning-Based Detection with Semantic Reasoning

Achim Reiz[a], Mohamad Albadawi[b], Kurt Sandkuhl[a], Matthias Vahl[b] and Dennis Sidin[b]

[a] *Rostock University, 18051 Rostock, Germany*
[b] *Fraunhofer IGD, Joachim-Jungius-Straße 11, 18059 Rostock, Germany*

Abstract

The company FutureTV produces and distributes self-produced videos in the fashion domain. It creates revenue through the placement of relevant advertising. The placement of apposite ads, though, requires an understanding of the contents of the videos. Until now, this tagging is created manually in a labor-intensive process. We believe that image recognition technologies can significantly decrease the need for manual involvement in the tagging process. However, the tagging of videos comes with additional challenges: Preliminary, new deep-learning models need to be trained on vast amounts of data obtained in a labor-intensive data-collection process. We suggest a new approach for the combining of deep-learning-based recognition with a semantic reasoning engine. Through the explicit declaration of knowledge fitting to the fashion categories present in the training data of the recognition system, we argue that it is possible to refine the recognition results and win extra knowledge beyond what is found in the neural net.

Keywords

Image Classification, Ontology, Semantic Augmentation, Deep Learning, Convolutional Neural Network, CNN

1. Introduction

Traditional, linear television is on a steady decline due to the rise of free and paid online content and video on demand. The increasing bandwidths combined with mobile flat rates, the possibility to create interaction with the users, and the uprising of new innovative media formats amplify this industry trend [2]. Especially, free online videos have a high reach in the advertising-relevant target group of 14 to 39-year-olds [3]. FutureTV, a content-marketing enterprise specialized in the creation and distribution of online-videos, fulfills the need for the rising demands of these short videos with self-produced, high-quality videos. These videos contain several scenes and show mostly fashion-related content. FutureTV creates monetary value through the placement of specific content-related advertising. For example, if the video shows a female face close-up wearing sunglasses and earrings, advertising should be placed for these specific items. Knowing the kind of fashion objects in the video has a direct impact on revenue and economic success. Historically, this approach heavily depended on the use of manual tagging. This approach is labor-intensive, costly, and challenging to scale up due to hiring and training the workforce. Automatic image detection technologies can reduce the need for workers and enable more efficient and cost-effective operation.

In the last eight years, the field of object recognition has been witnessing a revolution in terms of achieved recognition accuracy; mainly led by deep neural networks. However, that came with additional costs. These new smart models need costly computational power and vast amounts of costly annotated training data to deliver good performance. This data also needs to be balanced. That is, target objects

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.
EMAIL: achim.reiz@uni-rostock.de (A. 1); mohamad.albadawi@igd-r.fraunhofer.de (A. 2);
kurt.sandkuhl@uni-rostock.de (A. 3)); matthias.vahl@igd-r.fraunhofer.de (A. 4); dennis.sidin@igd-r.fraunhofer.de (A. 5)
ORCID: 0000-0003-1446-9670 (A. 1); 0000-0002-7431-8412 (A. 3); 0000-0003-1183-5823 (A. 4)
© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

need to be equally represented in the data. In a context where fine granular categories are highly desirable (e.g. FutureTV use case), it may be challenging to meet that condition of balance for objects that are in their nature not very common. In fashion categories, for instance, under the main category ‘coat’, a tailcoat is not as common as a raincoat.

The approach presented in this work aims at eliminating the need for annotated data corresponding to fine granular categories. By that, less overall data (and hence less effort) will be needed in the fine granular recognition task. Moreover, no imbalance problems will arise finding examples of the fine categories. The paper is structured as follows. The next section introduces the related work, followed by an overview of our new approach's technical architecture. Section four is then concerned with the evaluation of exemplary results. The approach is further motivated by an economic business case in section five. The paper concludes with an outlook on further research prospects.

2. Related Work

This related work section is structured in two parts. At first, we overview the used image recognition technologies and show the advantages of a hierarchical classification approach. In the next step, the state of the art regarding the connection of semantic technologies with deep learning is established.

2.1. Image Recognition Technologies

In recent years several neural networks architectures have been proposed for image classification and object detection. The ResNet [4] family represents one of the most widely used approaches for image classification. ResNet exhibits a simple but effective strategy of stacking large number of convolutional blocks that are furthermore coupled by shortcut identity connections; that enabled the building of increasingly deeper models leading to excellent performance. ResNeXt [5] is based on ResNet; it integrates a technique initially used by Inception [6] known as split- transform-merge. The input of an Inception module is split into lower-dimensional embeddings and then transformed by a set of filters before the results are merged and concatenated. Those aggregated transformations outperform the original ResNet modules even under the restricted condition of maintaining model and computational complexity. Most architectures are used in flat classifiers; some works [7–9] suggested using convolutional classifiers in a hierarchical fashion for better separation between visually similar objects. That way, a classifier's ability is concentrated toward such objects rather than being distributed among large number of objects categories in the flat approach.

Recently a competing family of neural networks called EfficientNets [10] has evolved. Remarkably, these architectures have not been handcrafted but discovered using neural architecture search. The aforementioned classification models are typically used as backbone models for modern object detection networks. These networks follow two main approaches: Two-stage architectures leveraging a proposal driven mechanism to generate a set of object locations that are then classified. One-stage detectors immediately regress bounding-box coordinates and classifications. Faster-RCNN [11] is a popular two-stage detector and is often used together with ResNet as the backbone model. It uses a sophisticated region proposal network in order to generate candidate object locations. One-stage detectors as SSD [12], YOLO [13], and RetinaNet [14] have been designed to be more efficient and thus faster than two-stage detector networks at the cost of an accuracy loss. Another approach to increase network efficiency has been the research of anchor-free network architectures as FCOS [15], which uses a fully convolutional network architecture. Recently a new family of one-stage architectures called EfficientDets [16] has been proposed, which uses the EfficientNets mentioned above as a backbone. This approach achieved similar accuracy with significantly fewer model parameters.

Over the last years, several scientific publications have researched classification and object detection in the context of fashion understanding and analysis [17–19]. Because of the excessive variety of clothing types and appearance and difficulties due to occlusion, classification, and detection of fashion objects remains a challenging problem. Several datasets like DeepFashion [12], DeepFashion2 [20], and ModaNet [21] of varying sizes have been made publicly available. However, depending on the specific type of fashion elements that need to be recognized, real-world use cases of these datasets are limited, and the synthesis of a suitable dataset is inevitable.

2.2. Connecting Semantic and Image Recognition

With the rising maturity of image recognition technologies, the connection of these technologies and semantic capabilities had seen some attention from the scientific community in the past years. Two literature reviews by Bhandari and Kulikajevas in 2018 [22] and by Ding et al. in 2019 [23] already collected the state of the art regarding the connection of semantic and deep learning technologies.

[22] first considers three different application areas for the interdisciplinary approach: The increasing accuracy in segmentation tasks, the automatic creation of image labels to annotate large libraries of former unstructured video and image content, and the recognition of part-of relationships of larger objects. Further, the paper names three domain-specific application scenarios: Robotics, to reduce the required computational capacity and to reduce the amount of detected false positives, geographic information science to translate the imagery into a GIS-ready format, and sports events to improve complex keyword searches.

[23] distincts between single object image recognition and multi-object image recognition. For the former, Ding et al. present examples for the connection of recognition algorithm with high-level semantics. This allows, e.g., the iterative detection of bird species on changing backgrounds or the more accurate classification of buildings. For the latter, the authors describe the opportunity to analyze the relationships between the targets in the multi-object environment for better analysis accuracy through the connection with WordNet or user-behavior identification in videos. Further examples for the connection of semantic and deep learning are the categorization and storing of information inside an ontology [25] or the rule-based indexing of CCTV [24].

On a high-level view, the related works share a similar core. From a detailed perspective, though, these approaches differ widely. The analysis of CCTV does not require a high granularity and detailed hierarchy for similar-looking objects. The connection of labels through WordNet does not function in a domain-specific environment with a specific business-related task.

Our work is concerned with fine-grained fashion categories, which need hard-to-find training data that may result in a highly imbalanced recognition problem. For that, we start with popular coarse fashion categories and leverage existing large-scale datasets to refine those. Our approach exploits the correlation between our coarse categories and the existing general-purpose large-scale recognition dataset. A semantic augmenter will be analyzing fashion elements in the light of knowledge extracted from the input imagery based on the Places dataset [25]. Back to the example '*coat'* from the previous subsection, it will be enough to detect a *coat* in the image; the semantic augmenter will take on from that point and infer a tailcoat knowing that the scene is a concert hall. There are currently no approaches that use ontologies to maximize the extracted information and reduce training costs of deep learning methodologies to the best of our knowledge.

3. Connecting semantic reasoning with image recognition

The project aims at developing an innovative approach by combining a technique from the symbolic and subsymbolic sub-disciplines of AI research. The aim is to apply knowledge captured in an ontology to improve object recognition in videos, based on an artificial neural network (ANN) and a deep-learning approach.

Figure 1 presents an overview of the architecture and orchestration of the newly created approach. The image recognition unit provides access to the ANN database containing available models. The models can either be for a single concept in the ontology or a combined model for several concepts.

The semantic augmentor unit provides access to the ontology. The ontology is supposed to capture the relevant knowledge for the application field of discovering fashion items in videos. There exists an ontological twin for every classification model in the deep learning part of the system. These twins are embedded in contextual knowledge like a taxonomy of fashion items, environments suitable for specific fashion categories (*mountain*, *skiing*, *outdoor*), social contexts relevant for fashion categories (*weddings*, *parties*), and more. See 3.2 for an extensive description of the ontology structure.

An upload of a video on the deep learning servers triggers an initial analysis, resulting in a scene (e.g., *airfield*, *bar*, *concert_hall*) classification, the detection of the gender (*male, female*), and the body area/fashion category. For each classified scene, a set of concepts from the ontology is considered relevant for this purpose. This "starter set" usually includes high-level concepts in the ontology (e.g., fashion category *TOP* for the gender *female* in the scene *outdoor*). Rather than searching for all potential fashion items, which results in many fashion items, the search starts with a subset of 2 to 8 objects. These results are returned to the analysis framework and forwarded to the semantic servers. The semantic service utilizes the shared taxonomy between the project partners and can infer contextualized knowledge fitting to the requested items. While the detection of a *bikini* or *bra* is likely in a swimming pool, the same item in a *hardware_store* or *ski_resort* is unlikely. The semantic service will, therefore, filter these elements.

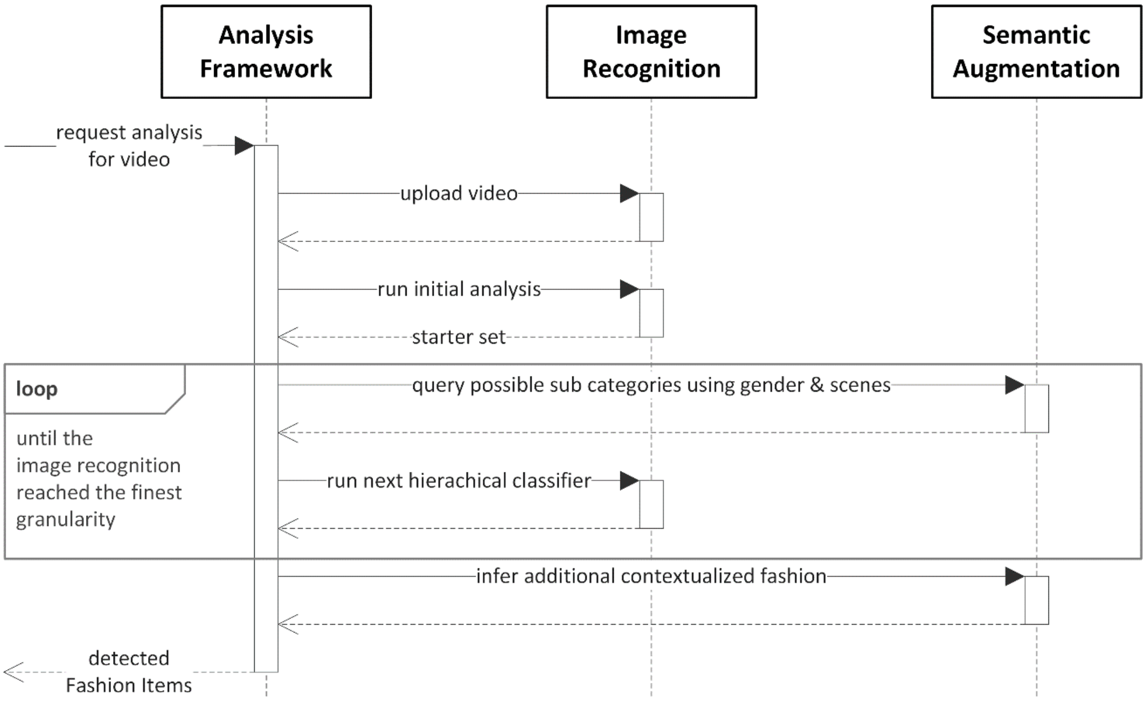
Analysis Framework
Image Recognition
Semantic Augmentation
request analysis for video
upload video
run initial analysis
starter set
loop
until the image recognition reached the finest granularity
query possible sub categories using gender & scenes
run next hierachical classifier
infer additional contextualized fashion
detected Fashion Items

Figure 1: Sequential Improvements of Detection Accuracy

The refined iteration can be triggered after a higher-level concept is detected, utilizing semantic reasoning based on the previous analysis. The iterative improvement process takes place on the common terminology that is defined for all project partners. Figure 2 shows an excerpt from this shared ontology. As an example, we could assume a medium shot of a woman in a swimming pool. The image recognition unit would recognize the scene *swimming_pool_indoor* and the fashion category *TOP*. The semantic component now can derive that no leaf item in the *TOPLAYER* and *MIDLAYER* category fits the classified situation and returns only the *LOWERLAYER* category as the next possible item in the given situation. For the next refinement iteration, two of the three available classifiers can be omitted in the image analysis, thus saving computational resources. With this approach, we reduce the effort required to object recognition compared to the brute force strategy of trying all existing concepts but expect to reach high tagging-quality. However, this expectation has to be validated in subsequent experiments.

Further, we can infer additional possible sub-categories of the detected fashion items through reasoning after reaching the image recognition service's finest search category. The shared ontology contains 62 classes. If the algorithm reaches a leaf element, no further iteration can be triggered at the image recognition-servers. To enable the deriving of possible, more accurate results beyond the image classifiers, the leaf elements of the shared ontology contain a link towards a more extensive, non-shared fashion vocabulary containing 693 elements. This larger ontology is based on the EU-funded fashion-brain project [26]. More information on the created fashion ontology, its evolution, and the underlying design decisions can be found in [27].

The integration of the more extensive fashion vocabulary with the semantic twins enables the inferring of new sub-classes without newly trained recognition classifiers through the use of semantic reasoning. The ontology stores matching sub-concepts for the detected classes and puts them into perspective to the already classified items, allowing the filtering of sub-concepts that might fit into a given situation. The latter is validated in the evaluation section.

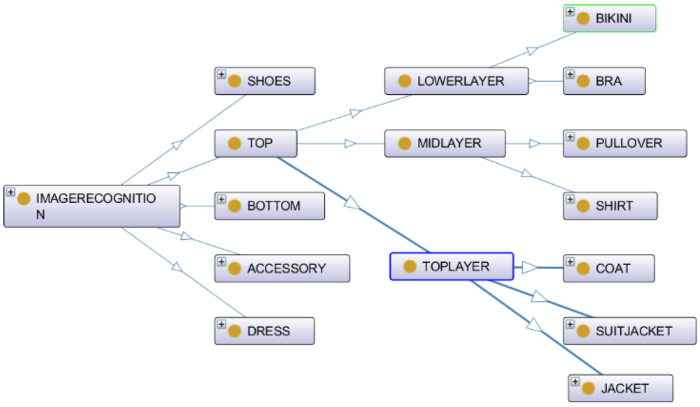

Figure 2: Excerpt of Shared Ontology

3.1. Structure of the Image Recognition Unit

The recognition unit consists of three main components that perform one of two main visual tasks: classification and detection. The components are next described in the same order of their use in the recognition unit. The first component is a ResNet18 classification model trained on the Places 365 dataset [25]; the model is trained and provided by the team responsible for Places. This network can distinguish between 365 scene categories covering a broad spectrum of environments seen in the real world. This component operates independently from the other two, which operate together to produce their results. The next component is an object detector that is trained to detect seven main types of objects, namely *Male_head, Female_head, Top, Bottom, Dress* (clothes covering the full body), *Accessory (hat, tie, bag …), and Shoes*. As for the detector's architecture, we leverage the more complex but more accurate ResNet50 as the backbone model for a Faster R-CNN detection model [11]. A dataset of 6000 images was prepared and annotated accordingly for the training of the model. The detector is trained only once and will then detect its seven categories regardless of which specific fashion element is present. It will always detect the region in an image representing a *top* object irrespective of which top is that *(T-shirt, jacket…)*. The detected objects are then to be further classified by the third component, which is a hierarchy of classifiers. We went for hierarchical classification because it offers the system a significant advantage compared to a flat classification approach; it delivers better accuracy with regard to objects with a similar appearance, as shown in works [7–9,28]. That is required in our use case as many fashion elements are very similar (e.g., *bra* vs. *bikini top*), and an appropriate differentiation is necessary for the semantic augmenter to work well. The enhanced accuracy comes with extra computational time; however, that is not a problem, as the system does not need to conform to any specific runtime requirements. Each classifier in the hierarchy corresponds to a non-leaf node in the tree of a shared fashion ontology with the semantic augmenter. Figure 2 shows an excerpt of the ontology. In total, 22 classifiers were trained; they can be controlled/run separately or as one entity.

Leaf nodes in that tree represent the final categories that the hierarchy can differentiate. In total, the tree has 37 leaf nodes corresponding to 37 fashion categories. These are the coarse fashion categories that the semantic augmenter can obtain from the recognition unit along with the gender information from the detector and the scene category from the first component for further refinement. The semantic augmenter also has access to the results of the intermediate classifiers in the hierarchy. To train the classifiers in our hierarchy, a dataset of around 60,000 images was prepared out of freely available images online. For each one of the 37 end-categories, about 1500 images were collected. For training intermediate classifiers, images from leaf categories were merged to represent parent categories. The images in this dataset were taken as tight around the objects as possible; that would yield training examples that imitate image areas under the bounding boxes delivered by the detector.

In the recognition unit, we followed a cloud-based computer vision as a service approach and provided our trained classification and detection models through a custom-built REST-API. This API allows the upload of a video or image file, selecting a trained neural network model for inference, and then the request of a classification or detection process. A feedback mechanism was also implemented in the recognition unit. Either the augmenter or a human user can mark-up false recognition results; the system can accumulate that information, and upon request, a training process can be triggered for arbitrary networks. The training process ends with the help of the early stopping stopping technique based on the validation loss.

Due to the fast-changing nature of computer vision research, we could not leverage the most recent model architectures like EfficinetDets as such models were not yet available at the time of implementing our solution. However, we developed our server and machine learning infrastructure with extensibility in mind. Therefore, it is possible to easily integrate new model architectures, and we are currently researching suitable candidates.

3.2. Structure of Semantic Image Augmentation Ontology

In this section, the semantic augmenting unit is described in more detail. The ontology is built in OWL and utilizes an extensive class structure without depending on individuals. Figure 3 presents the structure of the semantic augmentation ontology. In the following example, *JEANS_SHORTS* are detected by the image recognition unit. It is associated with the gender-class *female* and *male*. The fashion items are not connected directly to the scenes but through an intermediary object *occasion*. The ontology utilizes object properties for the connection of the various classes, modeled as subclass relationships. The ontology is publicly available at [29].

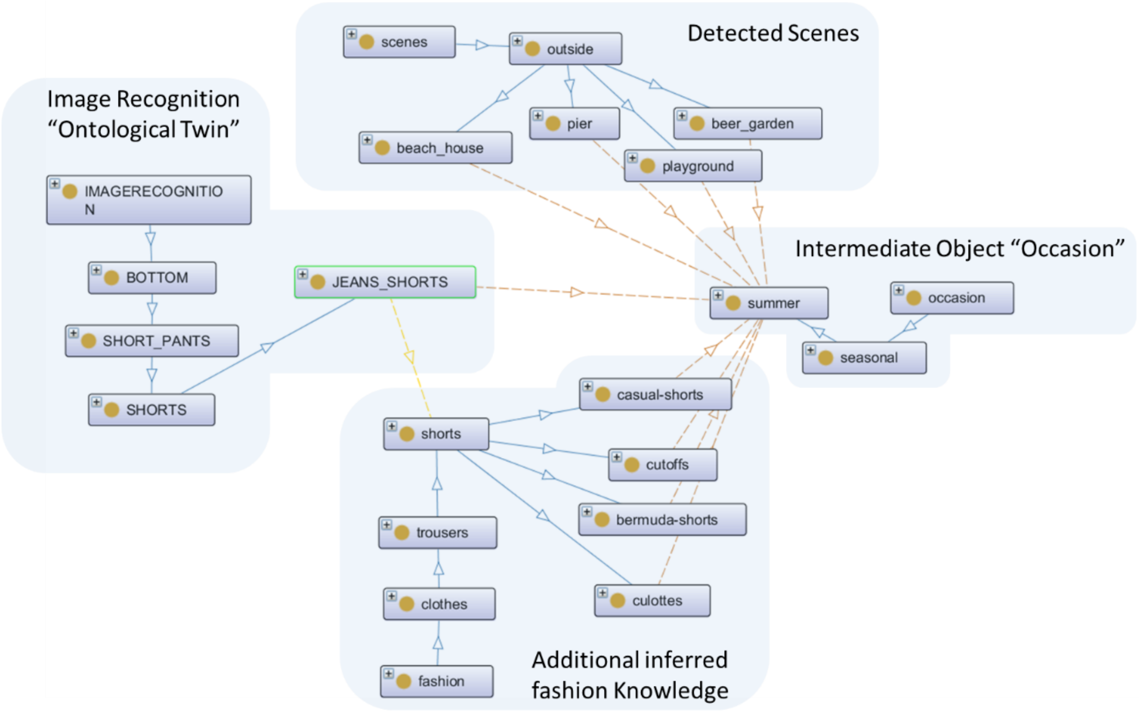

Figure 3: Example for the Encoded Additional Knowledge to a Detected *BUSINESS_TROUSERS* (Excerpt)

As the ontology consists of 365 recognizable scenes and over 750 fashion items, creating a connection between all of these items heavily increases the size of the ontology and the modeling effort. The in total 56 occasions reduce the complexity and ease the maintenance of the ontology. Taking the example of the JEANS_SHORTS, it allows to exclude *occasions*, among others, like *winter*, *formal-business*, or *high-class-events* with the associated, detectable *scenes* like *courthouse*, *ski_resort*, *office* or *dining_hall*.

Table 1
Accuracy of Image Recognition

Classifier	Classes	Top-1 Accuracy %
Accessory	Headcover, Eyewear, Bag, Tie	93.1
Headcover	Caps, Hats	90.4
Eyewear	Sunglasses, Clear glasses	91.0
Bag	Handbag, Backpack	91.8
Top	Lower layer, Mid layer, Top layer	90.5
Lower layer	Bikini, Bra	85.7
Mid layer	Pullover, Shirt	91.8
Top layer	Suit jacket, Coat, Jacket	90.8
Shirt	T-shirt, Business shirt, Auxiliary shirt	90.3
Dress Occasion	Formal dress, Casual dress	88.9
Dress Type	Long dress, Mid dress, Short dress	89.7
Shoes	Open shoes, Closed shoes	92.2
Open shoes	Open heels, Sandals, Flip-flops	90.6
Closed shoes	Boots, Flat shoes, Closed heels	91.9
Boots	Extra high boots, High boots	91.0
Flat shoes	Sneakers, Business shoes	91.5
Bottom	Short pants, Trousers, Skirt	93.0
Short pants	Shorts, Underwear	92.6
Trousers	Business trousers, Casual trousers	90.2
Skirt	Long skirt, Short skirt	89.9
Shorts	Jeans shorts, Other shorts	92.7
Casual trousers	Jeans trousers, Sport trousers, Others	91.1

Taking the example of Figure 3, the *IMAGERECOGNITION* class represents the iterative improvements of the image recognition service. For the example *JEANS_SHORTS*, at first, the body area is detected (*BOTTOM*), then the kind of trousers (In this case *TROUSERS*, other possibilities would be *SKIRT*, and *BUSINESS_PANTS*). At last, the clothing element itself is detected. At this point, the image-recognition classifier does not offer additional knowledge; it has reached the finest available granularity. The leaf element of the *IMAGERECOGNITION's* ontological twin now points to the more extensive fashion-knowledge base. As these elements are connected to the same describing attributes, the semantic engine can infer additional fashion items that fit the given situation using the same *gender* and *occasion/scenes* constrains. In our example, the larger, non-shared fashion ontology contains additional shorts-elements like *casual_shorts*, *cutoffs*, *bermuda-shorts*, *cuolottes*, and more (30 different kinds of shorts in total).

4. Evaluation of Image Recognition Unit

Improving on the accuracy of state-of-the-art machine learning models is out of the scope of this work. However, an evaluation of our models is important to make sure they are working properly. The ResNet18 classification model trained on Places365 was provide by the authors [25]. They provided multiple ResNet models with different number of layers. Accuracy figures for the 18-layers model were not mentioned. We evaluated the model on the test set of Places365 and reached a top-1 accuracy of 53.2% and a top-5 accuracy of 83.8%. Our 7-classes detection model was evaluated on a test set of 1000 images and reached an mAP of 51.7% using an intersection over union threshold of 0.5. Classifiers in the classification hierarchy were individually evaluated. Each classifier was assessed on a test set that has around 100 images per class. The results are summarized in Table 1.

Table 2
Results of the Semantic Augmentation for 12 Sample Images

#	Scenes	m/f	Image Detection	Semantic Augmentation
1	ocean, wave	f	pullover	NA
			other_shorts	NA
2	playground, corral	f	t-shirt	t-shirt: uni-t-shirt, pattern-t-shirt, print-t-shirt
			other_trousers	pants: Cargo-pants, casual-pants, chinos-and-khakis, corduroy, cropped-pants, dress-formal, joggers, knits, overalls, overall-pants, rain-pants
			sunglasses	sunglasses
3	ski_slope, snowfield	-	jacket	jackets: down-jackets, fleece-jackets, leather-jacket, puffers, shearling-jacket, sport-jackets, winter-jackets
			sport_trousers	sport-pants: snowboading pants; sport-jackets: snowboarding-jackets
4	stadium_soccer, stadium_baseball	m	caps	baseball-caps
			t-shirt	t-shirt
			other_shorts	shorts:sport-shorts: soccer-shorts
5	beauty_salon, dressing_room	f	bra	bras: strapless, sports-bras, push-up, minimizers, mastectom, demicup, convertible, bralettes
			underwear	underpants: bikinis, g-strings, hipsters, tangas, thongs, briefs
6	picnic_area, forest_path	f	short (dress)	baby-doll, jesery-dresses, summer-dresses, tank
			sneakers	sneakers
7	street, plaza	f	hats	hats: berets, ear-muffs, newsboy-caps, straw-hats, sun-hats, visors, fedoras
			axillary-shirt	t-shirt: uni-t-shirt, pattern-t-shirt, print-t-shirt, tank-tops
			long-skirt	skirts: mid-length-skirts, mini-skirts, long-skirts
			sandals	sandals: classical-sandals, flat-sandals, high-sandals
8	elevator_door, elevator_lobby	m	business-shirts	button-down-shirts, button-down-oxfords
			tie	ties-cummerbunds: bow-ties, neck-ties
9	office, home_office	m	t-shirt	t-shirts
			jeans_trousers	jeans: bootcut-jeans, skinny-jeans, slim-jeans, straight-leg-jeans, stretch-jeans
10	kitchen, wet_bar	f	axillary_shirt	t-shirts:pattern-t-shirt, print-t-shirt, uni-t-shirt, tank-tops
			mini_skirt	skirts: mid-length-skirts, mini-skirts, long-skirts
11	reception, beauty_salon	f	mid (dress)	jumper-dresses, sweater-dress, wrap, day-dresses, shirtdresses, knitted-dresses
			open_heels	boots, heels
12	pharmacy, bar	m	Business_Shirt	NA
			other_trousers	pants: cargo-pants, casual-pants, chinos-and-khakis, corduroy, corduroys, cropped-pants, dress-formal, joggers, knits, overalls, overalls-pants, wide-leg-pants

5. Evaluation of Semantic Augmentation

The current results of the new service look promising. For the evaluation, we choose a total of 12 different pictures. As the evaluation is focused on the performance of the semantic augmentation engine, we considered "perfect" image-detection results. Table 1 shows the detected scenes, the corresponding image detection items in their finest granularity, and the additional items inferred by the semantic augmentation engine. The last row, containing the inferred elements, has to be read the following:

If the semantic engine cannot infer any results, this is indicated by NA. Otherwise, the linked item is presented, followed by their fitting sub-items, if applicable. For example, in result #2, the image-recognition item *t-shirt* is linked with a *t-shirt* item in the more extensive fashion knowledge base. This linked item also has more detailed sub-items like *uni-t-shirt*, *pattern-t-shirt*, and *print-t-shirt*. The scenes constrain these subclasses. An example of this constraint can be found in #4. For the scenes *stadium_soccer* and *stadium_baseball*, there are no subitems of t-shirt linked as a proper fit.

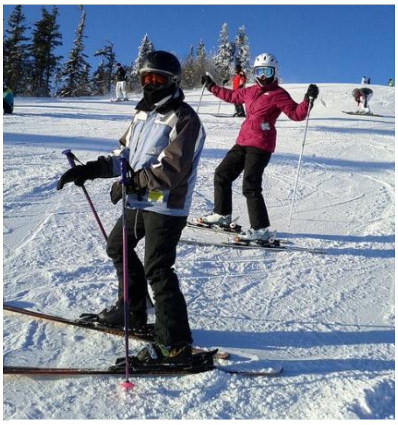

Figure 4: Man on the Ski-Slope by [1] (Item detection #3)

The motivation for the development of this prototype was the generation of advertisable fashion. Therefore, the focus was not the precise description of subclasses but the inferring of fashion-items that can be worn in a given situation and fit the elements detected by the image recognition technology. Taking example #3, showing a man on a ski-slope. The semantic linked in total seven different items for the detected jacket. Of these seven items, we deemed four of them fitting as advertisable in the given context *(down-jacket, winter-jacket, puffers, sport-jackets),* even though only two are exact sub-classes of specifying the presented object *(winter-jacket and sport-jacket).*

The 12 pictures contained a total of 28 fashion-objects detectable by the image recognition. These image recognition-objects are linked to 37 elements of the fashion knowledge base. The inclusion of sub-items extends the number of inferred items to 94. Of the 94 items, we rated 67 relevant for an advertising context, resulting in a precision value of 71,3%.

6. Expected Economic Benefits

The hybrid approach's economic potential can be illustrated by comparing the object recognition solely using machine learning with the combined semantic and deep learning (DL) approach. Let us assume that we have M videos and determine the relevant objects in video k. Furthermore, there are N classifiers available that have been trained in the DL approach, one for each object. What objects are relevant in what scene is further specified in the semantic net. To determine the relevant objects in k can be done only with the DL approach or with the combined semantic net and DL approach.

In the first case (without a semantic net), to identify all relevant objects in video k, all classifiers have to be applied because there is no information from the semantic net about what scenes exist, what objects characterize these situations and what additional objects are relevant in a scene. In this case, the effort E_k of identifying relevant objects for video k consists of using the approach with all available classifiers.

When using the semantic net, the classifiers of the DL approach first have to be used to identify the scene. The number of scenes is much smaller than the total number of objects for M, and the number of classifiers required to determine a specific scene also is much smaller than the total number of objects. In the second step, only the objects related to the identified scene are relevant, i.e., only this specified set of classifiers for the DL approach has to be used.

The effort required to determine the relevant objects in k consists of the effort for detecting the scene plus the effort for running the classifiers for the objects relevant in the detected situation. This is illustrated in Figure 5, where e_0 is the effort required to prepare scene recognition for video k on the initial analysis of the semantic net, e_i is the effort required to run an individual classifier, and N' is the number of required classifiers to determine all relevant objects. If for video k all objects (and available classifiers) are relevant, N is equal to N'. Then the effort of object recognition without the use of the semantic net will be less than with its use.

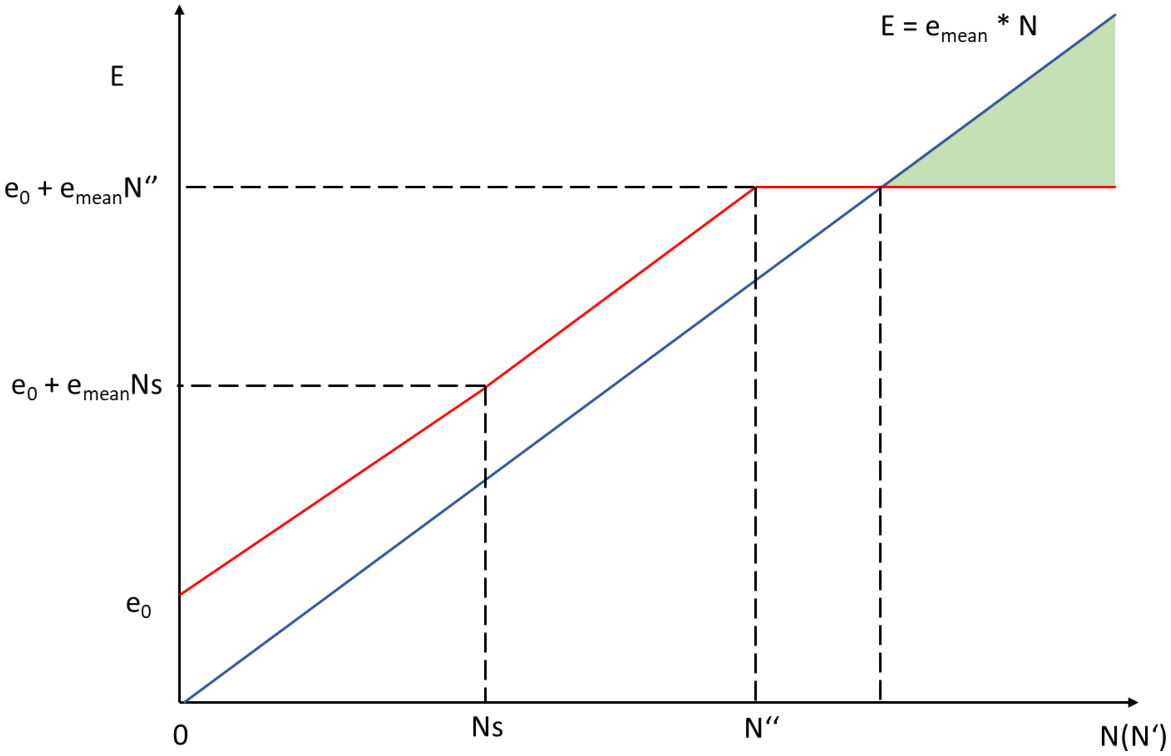

Figure 5: Comparing efforts required for object recognition with and without the semantic net.

Figure 5 shows the dependence graphs of the effort of object recognition on their number with and without using the semantic net. For descriptive purposes, the assumption has been made that the effort to run a particular classifier is the same, and it is e_{mean}. Then, the effort required for determining all relevant objects in k without using the semantic net is $E = e_{mean} * N$.

When using the semantic net for video k, an initial effort e_0 will be required to bootstrap the semantic net and select N' objects to be used for scene recognition. In this case, N' ∈ [1, N''], where N'' is the number of the applied classifiers for objects required for scene recognition, which usually does not exceed 5–6. It should be noted that the more distinctly the objects describe the relevant scenes, the smaller the value of N'' will be. When the situation can only be described by using many objects, the value of N'' will be high (i.e., a large number of classifiers must be used). If there is a set of objects significant for only one specific scene, the value of N'' will be low. Then, the required effort for object recognition with the help of the semantic net will be determined for N' < N'' according to the formula $E = e_0 + e_{mean}N'$, and for N' > N'' it will be determined by the formula $E = e0 + e_{mean}$ N''. Ns is the number of classifiers for determining a scene, and $E_k = e_0 + e_{mean}N'$ the required effort. The triangle colored in green is the expected benefit of the hybrid approach.

7. Discussion

The usage of automated object detection has the potential to replace the manual tagging of video contents and can, therefore, lead to significant monetary savings. However, the specific characteristics of the analysis of fashion-related videos present a challenge for implementing a classical object detection analysis. Due to the nature of moving pictures, many frames need to be analyzed and require enormous computing power. Some fashion objects look similar to each other and need additional context to be distinguished. In this paper, we proposed a combination of a deep-learning CNN through a hierarchical semantic network. We argue that this approach has the potential to lower the computational requirements and enhance precision. Furthermore, the extension of the novel detection requires less effort, and the deriving of additional information besides the image recognition data through semantically linked knowledge is possible.

In this work, we evaluated the semantic augmentation performance and showed how it could help bring us beyond what we can explicitly detect with typical recognition techniques. While the semantic augmenter and the image recognition system are now deployed online and the productive end-software is imminent (debugging phase), a full evaluation of the novel approach (image-recognition + semantic) is still pending. Therefore, this research endeavor's next steps are concerned with the numeric analysis of the characteristics in terms of computational requirements, end accuracy (recognition error + semantic error), training effort, and response times.

Acknowledgements

We like to thank FutureTV GmbH for the successful collaboration and are looking forward to working together on further innovative approaches relevant to the research and industry community. The results shown here are based on the joint project "Kl-based object recognition and semantic-supported content analysis (KOSlnA)" which is co-financed by the European Union from the European Regional Development Fund. Operational Programme Mecklenburg-Vorpommern 2014 to 2020 – Investments in growth and employment.

References

[1] I. Irving, Two Happy Ladies on top of Le Crete, 2012. https://flic.kr/p/dHrvi8 (accessed 30 November 2020).
[2] Goldmedia GmbH, Grugel Productions, WEB-TV-MONITOR 2019, 2019.
[3] SevenOne Media, View Time Report: Neue Perspektiven der Videonutzung, 2019.
[4] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, IEEE, Piscataway, NJ, 2016, pp. 770–778.
[5] H. Touvron, A. Vedaldi, M. Douze, H. Jégou, Fixing the train-test resolution discrepancy, 2019.
[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, IEEE, Piscataway, NJ, 2015, pp. 1–9.
[7] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, Y. Yu, HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition, in: 2015 IEEE International Conference on Computer Vision, Santiago, Chile, IEEE, Piscataway, NJ, 2015, pp. 2740–2748.
[8] C. Murdock, Z. Li, H. Zhou, T. Duerig, Blockout: Dynamic Model Selection for Hierarchical Deep Networks, in: 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, IEEE, Piscataway, NJ, 2016, pp. 2583–2591.
[9] T. Zhao, B. Zhang, M. He, W. Zhanga, N. Zhou, J. Yu, J. Fan, Embedding Visual Hierarchy with Deep Networks for Large-Scale Visual Recognition, IEEE Trans. Image Process. (2018). https://doi.org/10.1109/TIP.2018.2845118.
[10] M. Tan, Q. Le V, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, ArXiv abs/1905.11946 (2019).
[11] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.
[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single Shot MultiBox Detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer vision - ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11-14, 2016 proceedings, Springer, Cham, 2016, pp. 21–37.

[13] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, IEEE, Piscataway, NJ, 2016, pp. 779–788.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, in: 2017 IEEE International Conference on Computer Vision, Venice, IEEE, Piscataway, NJ, 2017, pp. 2999–3007.

[15] Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully Convolutional One-Stage Object Detection, in: Proceedings, 2019 International Conference on Computer Vision, Seoul, Korea (South), IEEE Computer Society, Conference Publishing Services, Los Alamitos, California, 2019, pp. 9626–9635.

[16] M. Tan, R. Pang, Q. Le V, EfficientDet: Scalable and Efficient Object Detection, in: CVPR 2020: Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[17] W. Yang, P. Luo, L. Lin, Clothing Co-parsing by Joint Image Segmentation and Labeling, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 3182–3189.

[18] J. Huang, R. Feris, Q. Chen, S. Yan, Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network, in: 2015 IEEE International Conference on Computer Vision, Santiago, Chile, IEEE, Piscataway, NJ, 2015, pp. 1062–1070.

[19] M.H. Kiapour, X. Han, S. Lazebnik, A.C. Berg, T.L. Berg, Where to Buy It: Matching Street Clothing Photos in Online Shops, in: 2015 IEEE International Conference on Computer Vision, Santiago, Chile, IEEE, Piscataway, NJ, 2015, pp. 3343–3351.

[20] Y. Ge, R. Zhang, X. Wang, X. Tang, P. Luo, DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, IEEE Computer Society, Los Alamitos, CA, 2019, pp. 5332–5340.

[21] S. Zheng, F. Yang, M.H. Kiapour, R. Piramuthu, ModaNet, in: Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, ACM, [Place of publication not identified], 2018, pp. 1670–1678.

[22] S. Bhandari, A. Kulikajevas, Ontology based image recognition: A review, CEUR Workshop Proceedings 2145 (2018).

[23] Z. Ding, L. Yao, B. Liu, J. Wu, Review of the Application of Ontology in the Field of Image Object Recognition, in: Proceedings of the 11th International Conference on Computer Modeling and Simulation - ICCMS 2019, North Rockhampton, QLD, Australia, ACM Press, New York, New York, USA, 2019, pp. 142–146.

[24] Alejandro Zambrano, Carlos Toro, Cesar Sanín, Edward Szczerbicki, Marcos Nieto, Ricardo Sotaquira, Video Semantic Analysis Framework based on Run-time Production Rules - Towards Cognitive Vision. https://doi.org/10.3217/jucs-021-06-0856.

[25] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 Million Image Database for Scene Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009.

[26] A. Checco, G. Demartini, A. Loeser, I. Arous, M. Khayati, M. Dantone, R. Koopmanschap, S. Stalinov, M. Kersten, Y. Zhang, FashionBrain Project: A Vision for Understanding Europe's Fashion Data Universe, 2017.

[27] A. Reiz, K. Sandkuhl, Design Decisions and Their Implications: An Ontology Quality Perspective, in: R.A. Buchmann, A. Polini, B. Johansson, D. Karagiannis (Eds.), Perspectives in Business Informatics Research, Springer International Publishing, Cham, 2020, pp. 111–127.

[28] X. Zhu, M. Bain, B-CNN: Branch Convolutional Neural Network for Hierarchical Classification, arXiv preprint arXiv:1709.09890 (2017).

[29] A. Reiz, Fashion-Ontology for the Connection of Semantic with Deep Learning, Rostock University, https://doi.org/10.5281/zenodo.4519359, 2021.