



Delft University of Technology

Rethinking the objectives of computer vision systems

Strafforello, O.

DOI

[10.4233/uuid:e08bbfbe-ecfd-481b-82f1-926564494436](https://doi.org/10.4233/uuid:e08bbfbe-ecfd-481b-82f1-926564494436)

Publication date

2024

Document Version

Final published version

Citation (APA)

Strafforello, O. (2024). *Rethinking the objectives of computer vision systems*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:e08bbfbe-ecfd-481b-82f1-926564494436>

Important note

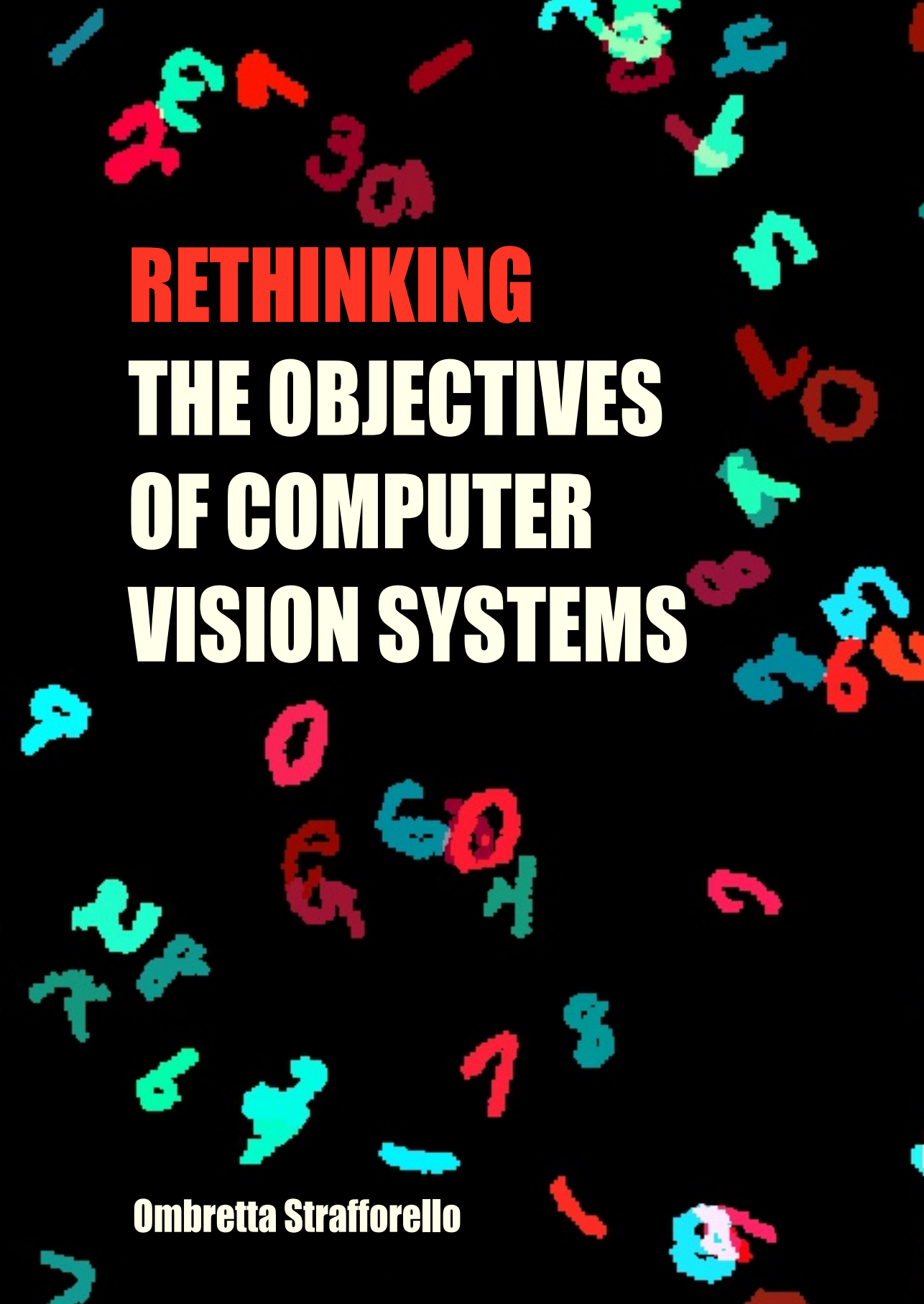
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



RETHINKING **THE OBJECTIVES** **OF COMPUTER** **VISION SYSTEMS**

Ombretta Strafforello

RETHINKING THE OBJECTIVES OF COMPUTER VISION SYSTEMS

RETHINKING THE OBJECTIVES OF COMPUTER VISION SYSTEMS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Thursday, 31 October 2024 at 12:30

by

Ombretta STRAFFORELLO

Master of Science in Computer Science,
Delft University of Technology, The Netherlands
born in Bordighera, Italy

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, <i>promotor</i>
Dr. J.C. van Gemert,	Delft University of Technology, <i>promotor</i>

Independent members:

Prof. dr. C.M. Jonker	Delft University of Technology
Dr. N. Strisciuglio	University of Twente
Prof. dr. P.H.N. de With	Eindhoven University of Technology
Dr. M.C. Popa	Maastricht University
Prof. dr. K.G. Langendoen	Delft University of Technology, reserve member

Other members:

Dr. K. Schutte,	TNO
-----------------	-----



Keywords: computer vision, object detection, human evaluation, action recognition, shortcut learning

Printed by: ProefschriftMaken.nl

Cover design: The cover is a creative reinterpretation of the MNIST dataset, originally created by Yann LeCun and colleagues, and widely used in computer vision research. The code used to generate the background design is available at <https://github.com/ombretta/color-mnist-cover>.

Copyright © 2024 by O. Strafforello

ISBN 978-94-6366-935-1

An electronic copy of this dissertation is available at
<https://repository.tudelft.nl/>.

CONTENTS

Summary	vii
Samenvatting	ix
Riassunto	xi
1 Introduction	1
1.1 Concepts of computer vision	3
1.2 Undesirable learning behaviors	13
1.3 Organization of this thesis	16
2 Humans disagree with the IoU for measuring object detector localization error	23
2.1 Introduction	23
2.2 Experimental approach	25
2.3 Results	27
2.4 Discussion	31
3 Aligning object detector bounding boxes with human preference	35
3.1 Related work	37
3.2 Do humans prefer larger detections?	39
3.3 Asymmetric regression loss to encourage larger detections	44
3.4 Conclusion	50
4 Long-term Behaviour Recognition in Videos with Actor-Focused Region Attention	57
4.1 Introduction	58
4.2 Related work	59
4.3 Method	60
4.4 Experiments	64
4.5 Conclusion	69
5 Are current long-term video understanding datasets long-term?	73
5.1 Introduction	74
5.2 Related work	77

5.3	Assessing long-term action recognition datasets	80
5.4	Results	81
5.5	Conclusion	87
6	Video BagNet: short temporal receptive fields increase robustness in long-term action recognition	93
6.1	Introduction	94
6.2	Related work	95
6.3	Method	98
6.4	Experiments	100
6.5	Conclusions	104
7	Discussion	111
7.1	Object detectors for user assistance	111
7.2	Long-term action recognition	113
7.3	Training and evaluating computer vision models	114
7.4	Final words	115
	Acknowledgments	121
	Curriculum Vitæ	127
	List of Publications	129

SUMMARY

Computer vision systems, such as image classifiers, object detectors and video analysis tools, serve diverse applications, ranging from autonomous vehicles and drone navigation to medical image analysis and anomaly inspection in the manufacturing industry. The development of these systems relies heavily on well established practices, which include the adoption of conventional training and evaluation metrics and benchmark datasets. However, we argue that standard approaches are sub-optimal with respect to the ultimate objectives of the computer vision systems. In this thesis, we question whether the training and evaluation of computer vision systems for object detection and long-term action recognition are typically aligned with human-defined end goals.

Object detectors are deployed for object tracking in autonomous vehicles and drones, but also as user assistive tools in medical image analysis and anomaly inspection in industry. Regardless of the end use, object detectors are trained with standard optimization and evaluation strategies. By investigating whether the optimization and evaluation methods of object detectors correlate with human quality judgments, we discover a discrepancy between established metrics and human preferences. To address this, we propose an alternative training loss that better aligns object detectors with human preference.

Subsequently, we ask whether object detections can be used to improve long-term human action recognition in videos. We find that explicitly focusing on the region containing the detected human is beneficial to long-term action recognition models. Unexpectedly, we also find that including a temporal attention module does not help recognizing the videos. Motivated by this result, we investigate how much temporal information is needed to solve long-term action recognition in three popular video datasets. Our results show that most of these videos can be recognized without any long-term temporal information. This suggests that models trained on these videos might exploit short-term shortcuts, instead of learning long-term temporal dependencies. Importantly, these models would not perform successfully on new videos where long-term reasoning is necessary.

As a follow-up, we investigate the impact of the temporal receptive field in long-term action recognition models. The size of the temporal receptive field determines the capability to encode long-term information in videos, like the actions order and duration. We experimentally verify that large temporal receptive fields

are sensitive to order and can overfit on the exact action orders seen at training time. Contrarily, short temporal receptive fields are more robust to order permutations and perform better on a current long-term video dataset. This result further demonstrates the irrelevance of long-term information in current long-term action recognition datasets. Our research findings highlight the importance of using training and evaluation metrics that match the intended use of the computer vision systems and choosing training and evaluation datasets that carefully represent the problem at hand.

SAMENVATTING

Computervisiesystemen, zoals beeldclassificatoren, objectdetectoren en videoanalysesoftware, dienen verschillende toepassingen, variërend van autonome voertuigen en drone-navigatie tot medische beeldanalyse en anomalie-inspectie in de productie-industrie. De ontwikkeling van deze systemen is sterk afhankelijk van gevestigde praktijken, waaronder het gebruik van conventionele trainings- en evaluatiemethoden en benchmarkdatasets. We betogen echter dat standaardbenaderingen suboptimaal zijn met betrekking tot de uiteindelijke doelstellingen van de computersystemen voor beeldanalyse. In dit proefschrift stellen we de vraag of de training en evaluatie van computersystemen voor objectdetectie en voor langdurige gebeurtenisherkenning (*long-term action recognition*) doorgaans overeenkomen met door mensen gedefinieerde einddoelen.

Objectdetectoren worden ingezet voor het volgen van objecten in autonome voertuigen en drones, maar ook als gebruikersondersteunende hulpmiddelen in medische beeldanalyse en anomalie-inspectie in de industrie. Ongeacht het eindgebruik worden objectdetectoren getraind met standaard optimalisatie- en evaluatiestrategieën. Door te onderzoeken of de optimalisatie- en evaluatiemethoden van objectdetectoren overeenkomen met menselijke kwaliteitsbeoordelingen, vinden we een discrepantie tussen gevestigde metrieken en menselijke voorkeuren. Om dit te adresseren, stellen we een alternatieve trainingsverliesfunctie (*training loss*) die ervoor zorgt dat objectdetectoren beter overeenkomen met menselijke voorkeur.

Vervolgens vragen we ons af of objectdetecties kunnen worden gebruikt om langdurige, menselijke gebeurtenisherkenning in video's te verbeteren. We vinden dat expliciet focussen op het gebied dat de gedetecteerde mens bevat gunstig is voor het modelleren van langdurige gebeurtenisherkenning. Onverwachts vinden we ook dat het opnemen van een temporele aandachtsmodule niet helpt bij het herkennen van de video's. Gemotiveerd door dit resultaat onderzoeken we hoeveel temporele informatie nodig is om langdurige gebeurtenisherkenning op te lossen in drie populaire videodatasets. Onze resultaten tonen aan dat de meerderheid van deze video's kunnen worden herkend zonder enige langdurige temporele informatie. Dit suggereert dat modellen die getraind zijn op deze video's mogelijk kortstondige *sluiproutes* benutten in plaats van langdurige temporele afhankelijkheden te leren. Belangrijk is dat deze modellen niet succesvol zouden presteren bij

nieuwe video's waar langetermijn redeneren nodig is.

Als vervolg hierop onderzoeken we de impact van het temporeel receptief veld in modellen voor langdurige gebeurtenisherkenning. De grootte van het temporeel receptief veld bepaalt het vermogen om langdurige informatie in video's te coderen, zoals de volgorde en duur van de gebeurtenissen. We verifiëren experimenteel dat grote temporeel receptieve velden gevoelig zijn voor volgorde en kunnen overfitten op de exacte volgorde van gebeurtenissen die geobserveerd zijn tijdens de training. Daarentegen zijn korte temporeel receptieve velden robuuster tegen volgordepermutaties en presteren ze beter op een huidige dataset van langdurige video's. Dit resultaat toont verder de irrelevantie aan van langdurige informatie in huidige datasets voor langdurige gebeurtenisherkenning. Onze onderzoeksresultaten benadrukken het belang van trainings- en evaluatiemetrieken gebruiken die overeenkomen met de beoogde toepassing van de computersystemen voor beeldanalyse, en het kiezen van trainings- en evaluatiedatasets die het voorliggende probleem zorgvuldig vertegenwoordigen.

RIASSUNTO

Sistemi di visione artificiale (*computer vision*), come classificatori di immagini, rilevatori di oggetti (*object detectors*) e strumenti di analisi video, servono a una vasta gamma di applicazioni, che vanno dalla guida autonoma dei veicoli alla navigazione dei droni fino all'analisi delle immagini mediche e all'ispezione delle anomalie nell'industria manifatturiera. Lo sviluppo di questi sistemi si basa pesantemente su pratiche consolidate, che includono l'adozione di metriche di *training* e valutazione convenzionali e set di dati di riferimento. Tuttavia, sosteniamo che questi approcci standard siano sub-ottimali rispetto agli obiettivi finali dei sistemi di computer vision. In questa tesi, mettiamo in discussione se il training e la valutazione dei sistemi di computer vision per il rilevamento degli oggetti e il riconoscimento delle azioni a lungo termine (*long-term action recognition*) siano tipicamente allineati con gli obiettivi finali definiti dall'uomo.

Gli object detectors vengono impiegati nei veicoli autonomi e nei droni per il tracciamento degli oggetti, ma anche come strumenti assistivi nell'analisi delle immagini mediche e nell'ispezione delle anomalie nell'industria. Indipendentemente dall'uso finale, i rilevatori di oggetti vengono addestrati con strategie standard di ottimizzazione e valutazione. Investigando se i metodi di ottimizzazione e valutazione dei rilevatori di oggetti correlino con i giudizi di qualità umana, scopriamo una discrepanza tra consolidate metriche di valutazione e le preferenze umane. Per affrontare questo problema, proponiamo una metrica di training alternativa che allinea meglio i rilevatori di oggetti con le preferenze umane.

Successivamente, ci chiediamo se gli object detectors possano essere utilizzati per migliorare il riconoscimento delle azioni umane a lungo termine nei video. I nostri risultati indicano che concentrarsi esplicitamente sulla regione contenente la persona che compie l'azione sia vantaggioso per i modelli di riconoscimento delle azioni a lungo termine. Inaspettatamente, scopriamo anche che includere un modulo di attenzione temporale non aiuta a riconoscere i video. Motivati da questo risultato, indaghiamo quanto sia necessaria l'informazione temporale per risolvere il riconoscimento delle azioni a lungo termine in tre popolari dataset di video. I nostri risultati mostrano che la maggior parte di questi video può essere riconosciuta senza alcuna informazione temporale a lungo termine. Ciò suggerisce che i modelli addestrati su questi video potrebbero sfruttare scorciatoie a breve termine, invece di apprendere dipendenze temporali a lungo termine. È importante

notare che questi modelli non avrebbero successo su nuovi video dove è necessario un ragionamento a lungo termine.

Come follow-up, indaghiamo l'impatto del campo recettivo temporale (*temporal receptive field*) nei modelli di riconoscimento delle azioni a lungo termine. La dimensione del campo recettivo temporale determina la capacità di codificare informazioni a lungo termine nei video, come l'ordine e la durata delle azioni. Verifichiamo sperimentalmente che ampi campi recettivi temporali sono sensibili all'ordine delle azioni. Questo può provocare *overfitting* agli ordini esatti delle azioni visti durante la fase di training. Al contrario, i campi recettivi temporali corti sono più robusti alle permutazioni dell'ordine e hanno prestazioni migliori su un dataset di video a lungo termine. Questo risultato dimostra ulteriormente l'irrelevanza delle informazioni a lungo termine negli attuali set di dati per il riconoscimento delle azioni a lungo termine. Le scoperte della nostra ricerca mettono in evidenza l'importanza di utilizzare metriche di addestramento e valutazione che corrispondano all'uso previsto dei sistemi di visione artificiale e di scegliere dataset di training e valutazione che rappresentino attentamente il problema in questione.

1

INTRODUCTION

Artificial Intelligence (AI) is present in our every-day life: smartphones use AI to categorize our pictures, we ask ChatGPT [1] to compose captivating poems and we generate beautiful visuals with Midjourney [2]. The AI technology that aims to understand and generate images and videos is called *computer vision*. Detecting objects in images and classifying human actions in videos are examples of tasks that computers and smartphones can solve automatically by means of computer vision algorithms. Relevant other applications can be found in a number of fields, including healthcare, manufacturing and autonomous driving.

Over the last twenty years, we have witnessed a tremendous progress in computer vision. In 2005, a state-of-the-art visual recognition systems could correctly categorize the images of the PASCAL VOC datasets in *four* classes: bicycles, cars, motorbikes, people [3]. Nowadays, computer vision systems can learn to recognize several *thousands* of image categories. A dataset that helped scale Computer Vision is ImageNet [4], first released in 2009. ImageNet is a collection of more than a million annotated images, belonging to one thousand classes. The dataset has been extended through the years and the current version contains more than 14M images. Because ImageNet is so large and diverse, successfully recognizing its images is considered a proxy to solve image classification. But is this really the case?

Let us inspect the ImageNet dataset a bit further. Each image has a class label that represents its content, the so called *ground-truth*. It is assumed that the ground-truth label is unique and that it thoroughly describes the image content. Figure 1.1 shows example of images from four different object classes. What is noticeable from this example is that, in addition to the ground-truth labels, the images contain objects that could be delineated by alternative labels. Namely, the ground-truth labels are sometimes insufficient or ambiguous. For example, the first image on the second row belongs to the class *Canoe*. However, the image also contains a person paddling and two Golden Retrievers.



Figure 1.1: Example of images from the ImageNet (ILSVRC2010) dataset [4]. The text in green shows the annotated ground-truth labels for four different classes. Alternative labels, marked in red, that fit the images just as well are considered wrong in the standard image classification training and evaluation paradigms.

In the standard image classification training and evaluation paradigms, classifying this image as *Golden Retriever* is simply considered wrong. As a consequence, a model that predicted the class *Golden Retriever* would be unfairly penalized. We might argue, instead, that an optimal image classifier should analyze the image thoroughly and predict the classes *Canoe*, *Golden Retriever*, *Man*, *Paddle*, *Water*. In 2021, Yun *et al.* [5] addressed this problem by relabeling ImageNet, taking into account multiple different objects in a single image. Nonetheless, the single labels of ImageNet are still widely used to evaluate state-of-the-art image recognition systems, testified by the 11.8k citations that the ImageNet paper obtained in 2023.

The ImageNet case is an example where the method used to develop computer vision systems – here, *training and evaluating algorithms to classify the single-label images of ImageNet* – do not exhaustively serve the intended objective: *automatic visual recognition*. In this thesis, we question whether there exist other cases where the development of computer vision solutions does not align with the intended objective set by humans. In particular, in this dissertation, we focus on two applications: object detectors as user assistive tools and long-term action recognition in videos. Our findings reveal that: 1) There exist a mismatch between the user preference and the evaluation metrics commonly used to evaluate object detectors; 2) Current long-term action recognition datasets do not encourage learning long-term reasoning in computer vision models, but rather the use of unintended shortcuts.

1.1 CONCEPTS OF COMPUTER VISION

"Human beings live in the realm of nature, they are constantly surrounded by it and interact with it." Dialectical Materialism (A. Spirkin).

Since the early months of life, humans learn to interact with the world. We do so by receiving sensory stimuli and reacting to them. Among the five senses, *vision* collects a large source of information that humans use to navigate in the world. A fundamental question in AI is: Can we teach *machines* to understand our visual word like humans do?

Computer vision is the science that tries to make sense of the visual world from image sensory data, like monocular and binocular cameras. This thesis focuses on understanding images and videos captured by standard digital cameras. This section introduces the computer vision tasks and techniques that will be the subject of the upcoming chapters.

1.1.1 IMAGE CLASSIFICATION

Image classification is a core task in computer vision. The goal is to predict the label that best describes the subject in a given image. The classification can be *single-label*, if one main object is present in the image, or *multi-label*, if the image contains multiple objects. In computer vision, image classification is achieved by training a machine learning model to *learn* a function that maps an image to its class label(s). During the training process, the model learns to classify a dataset of labeled images whose true class labels, known as the "ground-truth", are provided. A successfully trained model should be able to correctly classify new, previously unseen images.

Typically, the model input comprises gray-scale pixel values of RGB values, ranging from 0 to 255. In the RGB representation, which is a triplet of values describing *red*, *green* and *blue*, the input is said to have three *channels*. The model outputs a vector of scores, indicating how likely the input image is to belong to a predetermined set of classes.

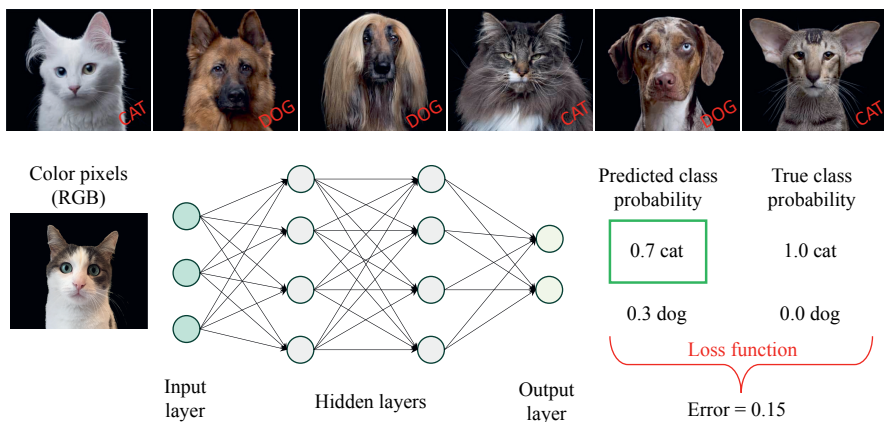


Figure 1.2: Top: Example of single-label image classification. Is there a cat or a dog in the images? The text in red shows the *ground-truth* label. Bottom: Illustration of a model for image classification. The model predicts class probabilities, which are compared to the ground-truth. The error is measured by the loss function and is minimized during training.

The state-of-the-art models used in computer vision are called *neural networks*. The name comes from the fact that these models are loosely inspired by the neu-

ral activity in the brain. Neural networks transform the input image representation through a set of sequential operations, consisting of linear functions that have learnable parameters and non-linear activation functions. A set of linear transformation and activation functions composes one model *layer*. Current models typically contain several layers: that is why they are said to be *deep* neural networks. This structure makes deep neural networks potentially capable to approximate complex mathematical functions.

Neural networks for image classification can flexibly learn which visual features to use to make a good classification. To distinguish cats and dogs, the model could learn to consider, among others, the length of the ears, the shape of the muzzle, the roundness of the pupils, the size of the nose and the texture of the fur. However, as visible in Figure 1.2, these features can vary significantly across samples of the same class. In addition, the same feature might appear differently with changes in light conditions and camera position. To guarantee that neural networks learn features that transfer robustly to new images of cats and dogs, the number of images seen at training time should be vast and should capture as many as possible variations of cats and dogs. In addition, neural networks should have enough *capacity* to store all features variations, namely, a large number of learnable parameters.

During training, the model parameters are repeatedly updated to minimize the amount of misclassified images. The misclassification error is measured by a differentiable *loss function*. In single-label classification, a commonly used loss function is Cross-Entropy (CE):

$$L_{CE} = - \sum_{i=1}^n t_i \log p_i, \quad (1.1)$$

where n is the number of possible classes, t is the true probability of the image corresponding to class i , namely 0 or 1, and p_i is the model prediction. If the model learned to output $p_i = 1$ for every class, L_{CE} would always be zero. To prevent this trivial solution, the softmax activation function is applied on the model output. This ensures that the predicted scores range from 0 to 1 and sum to 1, like the class probability. This way, L_{CE} approaches infinity when the wrong class is predicted.

In multi-label classification, multiple classes can occur simultaneously in one image. For each co-occurring class, the true class probability t is equal to 1. This scenario might be useful if a cat and a dog appear in the same image and we want to identify both. In this case, the softmax activation function is replaced by a sigmoid, to allow the sum of the predicted scores to be greater than 1, and the Binary Cross-Entropy loss is used:

$$L_{BCE} = - \sum_{i=1}^n t_i \log p_i + (1 - t_i) \log(1 - p_i). \quad (1.2)$$

L_{BCE} is minimized when the predicted score of each object class contained in the image is 1 and the predicted score for the other classes is 0.

The algorithm used to optimize neural networks is called *gradient descent*. Repeatedly, the model parameters are updated to move towards the (local) minimum of the loss function. The direction of this update step is given by the gradient of the loss function with respect to the model parameters. Many versions of this optimization algorithm have been proposed to improve convergence speed and stability. Two popular optimizers are Stochastic Gradient Descent [6] and Adam [7].

To assess the performance of an image classifier, the trained models are tested on a separate dataset, called the *test set*. A commonly used evaluation metric is the percentage of correctly classified images. However, this number can be misleading if the dataset is unbalanced. For example, if the test set contained 990 images of dogs and only 10 image of cats, a faulty model that always predicts the class *dog* would show 99% accuracy. In this case, *precision* and *recall* are more informative metrics:

$$\text{Precision} = \frac{\# \text{ true positives}}{\# \text{ true positives} + \# \text{ false positives}}, \quad (1.3)$$

$$\text{Recall} = \frac{\# \text{ true positives}}{\# \text{ true positives} + \# \text{ false negatives}}. \quad (1.4)$$

The faulty model would result in perfect precision for the class *cat* (no pictures of dogs have been wrongly classified as "cat"), but zero recall (no cats have been correctly identified). Often, precision and recall are combined into a single metric, called *F1 Score*, that corresponds to their harmonic mean:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (1.5)$$

In multi-label classification, image classifiers predict scores ranging from 0 to 1 for each class in the dataset. Class scores that are above a pre-set confidence threshold are treated as positives. High thresholds encourage high precision, because they allow for positive predictions only when the model is very confident. On the other hand, lower thresholds allow for higher recall. In an ideal model, lowering the confidence threshold increases the recall without drastically compromising the precision. A multi-label classification model can be evaluated by calculating the F1 Score for each class or from its *precision-recall curve*. This curve can be constructed by plotting the values for the precision and recall obtained with varying confidence thresholds on, respectively, the y and x-axis. The information given by the precision-recall curve can be summarized by two metrics: the *Area Under the Curve* (AUC) or the *Average Precision* (AP). The AP is calculated by taking the weighted mean of the precision achieved at each confidence threshold, with the

increase in recall from the previous threshold used as the weight. For n thresholds,

$$AP = \sum_{i=1}^n (R_i - R_{i-1})P_i, \quad (1.6)$$

where P_i and R_i are the precision and recall obtained for the i -th confidence threshold. The AP is calculated for each class and the *mean Average Precision* (mAP) is usually reported.

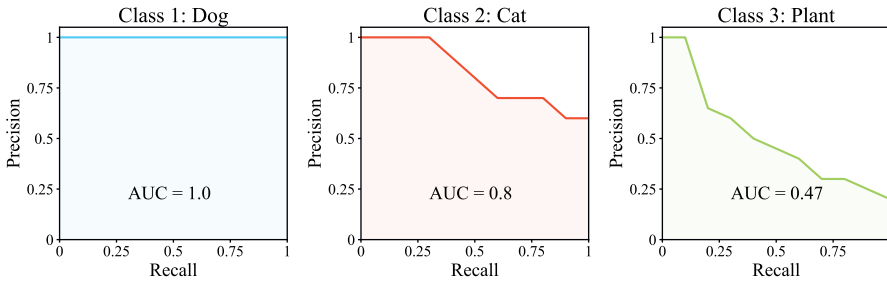


Figure 1.3: Illustration of the precision-recall curves for three classes. The classification is perfect for the class *dog*, good for class *cat* and poor for class *plant*.

Several neural network designs for image classification have been proposed over the years. The *multilayer perceptron* (MLP) is the simplest model. It consists of at least three layers of linear transformations and non-linear activation functions. MLPs are termed *fully-connected* because, in each layer, every unit in the input vector influences every unit in the output vector. To make 2D images compatible with the 1D input vectors expected by MLPs, images are first flattened. This solution is not ideal, as it disregards the information about the spatial disposition within the image. Another drawback is the sensitivity of MLPs to translations within an image, meaning that even a slight shift in object position can drastically alter the network's representation. As a consequence, the predicted class for the same object in two different positions might be different.

Convolutional neural networks (CNNs), illustrated in Figure 1.4, solve the limitations of MLPs by leveraging the *convolution* operation. In a convolutional layer, the image representation is created by comparing the input image with a set of small 2D kernels, shifting across height and width. Each convolutional kernel is a matrix of parameters optimized to encode a specific 2D pattern. The similarity between an image region and a convolutional kernel is measured through the dot product, which is high when the two strongly correlate. Since multiple kernels are convolved with the input, the output features include multiple *channels*,

each encoding diverse information. The CNN architecture preserves information relative to the spatial structure of the image. In addition, because of the shifting, the same convolutional kernel can efficiently detect the same pattern in different locations. Usually, in a convolutional layer, the convolution operation is followed by a ReLU activation function and pooling. Pooling reduces the dimension of the feature maps to reduce memory and time complexity. This spatial compression also cause the subsequent layers to encode patterns at a larger scale. Specifically, the size of the image region that influences the output in each convolutional layer is called *receptive field*. The growth of the receptive field through the layers allows the model to capture multiple levels of abstraction in the image, from edges and basic textures to complex geometrical shapes.

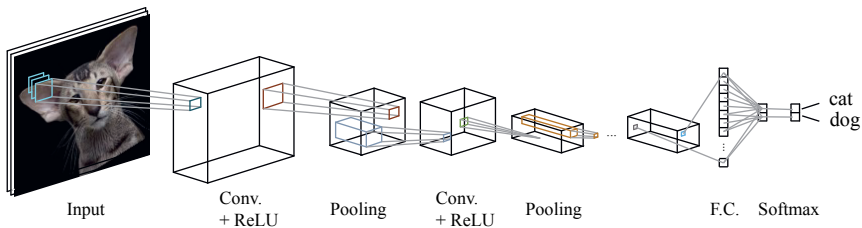


Figure 1.4: Illustration of a convolutional neural network for classification of *cat* vs. *dog*. The model contains several convolutional layers, comprising the convolution operation, a ReLU activation function and pooling. In the last layer, the image representation is flattened in a 1D vector, input to a fully connected layer (F.C.) and a softmax to predict scores per class.

The pioneering CNN architecture, AlexNet [8], was introduced in 2012. Since then, CNNs have been a go-to model for image classification and many variants were proposed. Notably, the residual neural networks (*ResNet*) [9] enhanced image classification through an architecture change that enables the successful training of very deep models. Only recently, an alternative model inspired by natural language processing, the *vision transformer* [10], has become a popular competitor of convolutional neural networks. One of the main difference of transformers is the *global* receptive field, which can capture spatial relationships anywhere across the image at any layer. This guarantees a higher level flexibility compared to CNNs. However, despite the competitive results achieved by vision transformers in the last three years, there is yet no evidence that these models are superior to convolutional neural networks [11]. Besides image classification, convolution is deployed for feature extraction in many computer vision tasks, including object detection

and action recognition.

1.1.2 OBJECT DETECTION

The task of object detection consists in predicting the class *and* the location of every object in a given image. Object detectors find application in several fields, including anomaly detection in medical images, like MRIs, CT scans and x-rays, crop monitoring and pest detection in agriculture, and manufacturing, by scanning products on assembly lines. The localization result is usually represented through a rectangular *bounding box* centered around each object of interest. Compared to image classification, object detection has an additional level of complexity which results in the need of sophisticated model architectures. An object detector takes as input image color pixels and, for each object, outputs class probability scores and box coordinates, usually expressed by the coordinates (x,y) of the top left box vertex and the height and width of the box.

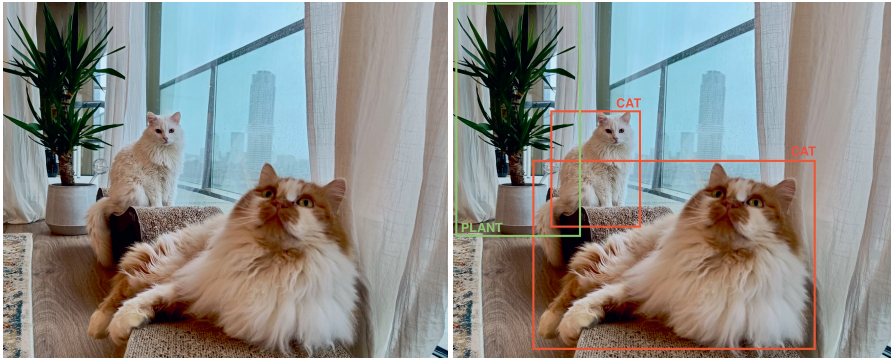


Figure 1.5: An object detector's input (left) and output (right).

Training an object detector is performed in a similar way as for an image classifier, by minimizing a loss function. Different from image classification, the object detection loss comprises two terms: the classification error, usually given by the Cross-Entropy loss, and a localization error. This is expressed by a regression loss, like L1 and L2, or the Smooth L1 loss, which generally provides a more stable convergence. Given the ground-truth $t = \{t_x, t_y, t_h, t_w\}$ and the prediction $p = \{p_x, p_y, p_h, p_w\}$, expressing the box coordinates of the top-left corner (x,y), height and width (h,w):

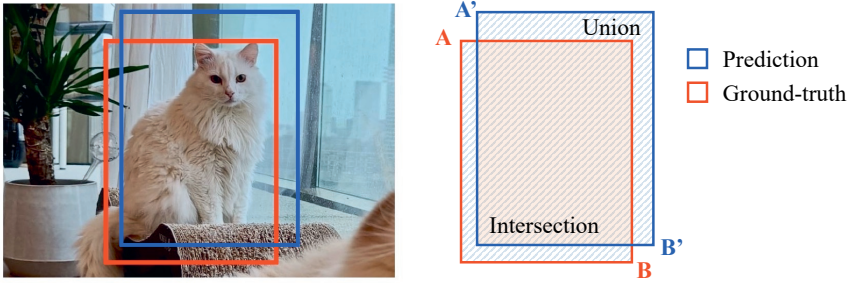
$$L1 = \sum_{i \in \{x, y, h, w\}} |t_i - p_i|; \quad (1.7)$$

$$L2 = \sum_{i \in \{x, y, h, w\}} (t_i - p_i)^2; \quad (1.8)$$

$$L1_{Smooth} = \sum_{i \in \{x, y, h, w\}} l_i, \quad l_i = \begin{cases} \frac{1}{2\beta}(t_i - p_i)^2, & \text{if } |t_i - p_i| < \beta, \\ |t_i - p_i| - \frac{\beta}{2}, & \text{otherwise.} \end{cases} \quad (1.9)$$

In the Smooth L1 loss, β is a hyperparameter that determines the smoothness of the loss function around the origin.

The localization accuracy is typically measured by means of *Intersection over Union* (IoU), which indicates how much the predicted box and the ground-truth box overlap. Boxes with different positions and size might result in the same amount of overlap with the ground-truth box. Therefore, the IoU metric is to some degree not sensitive to box translations and box size. A detection is considered correct if the predicted object class is correct and the IoU is above the acceptance threshold. Similarly to multi-label classification, the performance of object detectors is expressed in terms of mean Average Precision (mAP). Usually, the value of the mAP for different IoU thresholds (e.g., from 0.5 to 0.95) is reported.



Given $A = (475, 1082)$, $B = (1255, 70)$, $A' = (542, 1223)$, $B' = (1350, 148)$:

$$\text{IoU} = \frac{\text{Area}(\text{Intersection})}{\text{Area}(\text{Union})} = \frac{(x_B - x_{A'}) * (y_A - y_{B'})}{\text{Area}(\text{GT}) + \text{Area}(\text{Pred.}) - (x_B - x_{A'}) * (y_A - y_{B'})} = \frac{754672}{903288} \approx 0.835$$

Figure 1.6: Illustration of the calculation of Intersection over Union (IoU) score.

The model architectures for object detection can be grouped in two main categories: two-stage detectors and one-stage detectors. Both methods utilize models for image feature extraction, like CNNs or vision transformers. In *two-stage* detectors, region of interest are first extracted from the image through an algorithm called *selective search* [12]. The localization and classification of an object is then performed for each region of interest. Fast R-CNN [13] is a pioneering approach

belonging to this category that, by design, does not require storing in memory the proposed image regions. Among the advantages of this approach is the possibility of sharing of information between the classification and localization modules, which improves efficiency and accuracy.

One-stage detectors, like YOLO [14], perform the detection task without relying on a preliminary region proposal step. These models predict multiple boxes and class probabilities simultaneously from the entire input image. While two-stage detectors are usually more accurate, one-stage detectors are significantly faster, thus preferable for applications where inference speed is more important than accuracy.

Several datasets have been proposed to train object detections systems, notably the MS COCO [15] dataset. MS COCO comprises 91 objects classes appearing in 328k labeled images. The images contain multiple objects, possibly cluttered and occluded, which makes the detection challenging. The dataset was annotated through crowdsourcing. Specifically, crowdworkers were instructed to identify object instances in a given image and, in a second stage, manually draw the outline of each object. Rectangular bounding boxes were automatically extracted from the outlines. Consequently, MS COCO's bounding boxes are tight around each object. Because of its large size and the variety of the images, MS COCO effectively captures the complexities of object detection in real-world settings. Since its release, MS COCO has been extensively used and cited by over 41k research papers.

All the proposed object detectors evaluated on this dataset are considered successful if they predict bounding boxes as tight as possible to the objects of interest. This approach is standard in the literature, but it disregards the specific end use of the object detectors. In this thesis, we focus on the case when object detections are shown as an end product to humans. Given this setup, we investigate whether predicting tight bounding boxes is always in line with human preference. We also question whether the IoU metric, which is to some extent invariant to the box position and size, reflects well the quality judgments of humans.

1.1.3 ACTION RECOGNITION

Another task in computer vision is action recognition, which consists in predicting a label that best describes what is happening in a given video. Applications of action recognition can be found in surveillance, sport analytics or industrial automation. A baseline method for action recognition is using an image classifier on individual video frames. However, this approach cannot model temporal information, which is necessary to distinguish certain types of actions, like *opening a door* vs. *closing a door*. On the other hand, analyzing the video data altogether provides

richer information, including temporal dynamics, motion and speed.

Video data can be represented as a spatio-temporal 3D volume, made of 2D frames stacked along the time dimension. One of the popular solutions to model spatio-temporal data is 3D convolution. Its functioning is analogous to 2D convolution, except for the input, convolutional kernel and output being 3D volumes. The 3D kernels capture spatial and temporal information simultaneously, by shifting through the height, width and time dimensions. In a 3D convolutional network (3D CNN), a layer comprises 3D convolution, non-linear activation functions and 3D pooling. Thanks to the 3D convolution and 3D pooling, the model receptive fields grows both in space and time, making it possible to capture information over a large time-span in the deeper layers. Popular 3D CNNs architectures are I3D [16] and 3D ResNet [17], whose temporal receptive fields in the last layer measure, respectively, 99 and 217 frames.

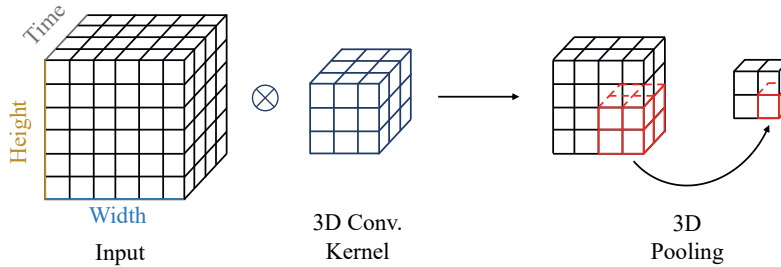


Figure 1.7: Illustration of 3D convolution (denoted by \otimes) and pooling operations, used to extract information in spatio-temporal input data.

It is a common practice to pre-train action recognition models on large-scale video datasets, like Kinetics [16], to learn various spatio-temporal patterns that might be useful to recognize actions in smaller downstream datasets. However, Byvshev *et al.* [18] discovered that Kinetics is biased towards appearance and can be largely solved without temporal information. As a consequence, models trained on this dataset might learn to focus on static information, without effectively encoding motion and other temporal dynamics. Other pre-training datasets, like HowTo1M [19], should be chosen for tasks that require action recognition models to understand temporal information.

In action recognition, the input videos are limited to a few seconds duration and are trimmed to contain a single action. This simple approach might not be representative of a video recorded in the wild. As an example, a video recording of a soccer game usually lasts several minutes and contains multiple actions, like *kick-*

ing, running, hitting the ball. Consecutive actions might be correlated and form a *long-term action*, in this case, *playing a soccer game*. To study *long-term action recognition*, several datasets and methods have been proposed. Current datasets mostly involve cooking, sports or instructional videos. Since this type of video is longer in duration and is made of multiple short actions over time, models with large temporal receptive field have been designed to capture long-term dependencies over a larger time-span than traditional 3D CNNs. Some examples include multi-scale temporal convolution [20] or modeling long videos with graph structures [21]. Inspired by the appearance bias discovered by Byvshev *et al.*, in this thesis, we challenge the assumption that long-term information is necessary to solve current long-term video datasets and study the performance of models with limited temporal receptive field.

1.2 UNDESIRABLE LEARNING BEHAVIORS

Deep neural networks, deployed in computer vision, are typically trained to achieve a specific objective, such as accurate image classification or action recognition. However, the learning process of these models can sometimes lead to sub-optimal results. In this thesis, we investigate whether the learning behavior of computer vision models aligns with the final objective. This section illustrates two undesired learning behaviors that are common in deep neural networks, namely *overfitting* and *shortcut learning*.

1.2.1 OVERFITTING

Computer vision models learn to perform a task on the training dataset. If the training process is successful, hopefully the models should be able to perform the task on new data. However, in situations where the training data is limited and the models have large capacity, the models might memorize the specific characteristics of the training data, including random noise, and fail to generalize on new data. This phenomenon is called *overfitting* and is a common undesired behavior in neural networks. To evaluate the models performance on new data, it is common practice to split the dataset in a training set and a test set. Since the two sets come from the same data distribution, they are independent and identically distributed (i.i.d.). Observing perfect training accuracy and poor accuracy on the i.i.d. test set signals the presence of overfitting.

It is possible to mitigate overfitting deploying various methods. One of these is data augmentation, which consists in artificially increasing the amount of training

data by creating modified copies of the original data samples. In images, common modifications include random cropping, horizontal flipping, affine transformations and color jittering. A second approach is by adding a regularization term to the loss function, which inhibits the model from learning too complex solution, thereby preventing memorizing overly specific characteristics of the training data.

1.2.2 SHORTCUT LEARNING

Shortcut learning is an undesirable learning behavior that can be encountered in deep neural networks. It occurs when a model learns a decision rule that successfully solves a task not by reasoning in a human-like fashion, but by means of unintended cues. Two noteworthy examples of shortcut learning have been found in image classification. Geirhos *et al.* [22] showed that convolutional neural networks trained on ImageNet are prone to focus on object texture over shape. Despite showing high accuracy on ImageNet, these models struggle to recognize line drawings and silhouettes, a task that humans tend to perform easily. In addition, the BagNet model proposed by Brendel *et al.* [23] revealed the use of unintended features to recognize some of the ImageNet classes, for instance using the fingers of fishermen to recognize *Tench*, a cyprinid fish.

Compared to overfitting, shortcut learning is harder to discover. While overfitting manifests itself when the test accuracy is significantly lower than the training accuracy, even with i.i.d. data, in shortcut learning both training and test accuracy can be high. Therefore, shortcut learning can be diagnosed only by testing on out of distribution (o.o.d.) samples or carefully analyzing what is causing the model predictions. Geirhos *et al.* illustrated these different learning behaviors [24] in a taxonomy, shown in Figure 1.8.

While overfitting is attributable to inadequate training, shortcut learning might be due to specific characteristics of the available data. For example, rosette fur texture is a discriminative feature of the class *Leopard*. In the absence of other object classes containing leopard fur texture, an image classification model can learn to predict *Leopard* solely by detecting this specific texture, while ignoring the rest of the image. This prediction rule can lead to the misclassification of images containing leopard print clothing. Analogously, static cues in videos might be used to recognize actions, overlooking temporal dynamics. For instance, action recognition models could overly exploit the correlation between actions and objects observed during training [25]. If the object *piano* appears only in videos belonging to the action class *Playing Piano*, an action recognition model might predict this class every time a piano appears in a video [26]. The bias towards appearance in the Kinetics video dataset, analyzed by Byvshev *et al.* [18], can be seen as a shortcut opportu-

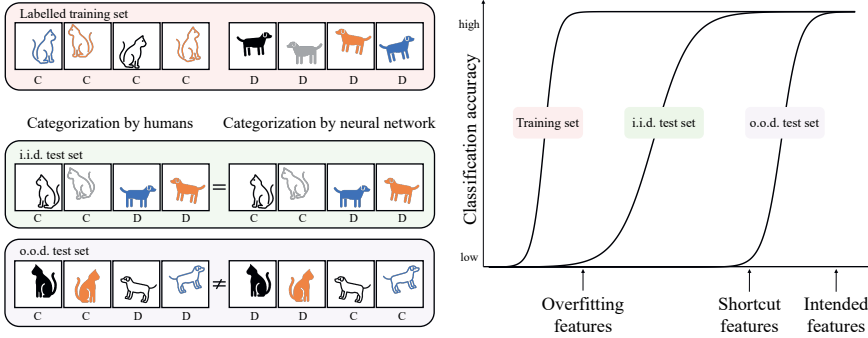


Figure 1.8: Illustration of the shortcut learning problem in neural networks, inspired by Geirhos et al. [24]. Left: A toy classification problem, cats (C) vs. dogs (D). Humans consistently classify the images based on the characteristic shape of cats and dogs, while a model using shortcut features learns to distinguish the images based on the shape color fill. Right: Taxonomy of decision rules. Using overfitting features results in poor classification on any dataset different than the training set, whereas shortcut features might fail only on out of distribution test sets. Human-like intended features perform successfully on any dataset.

nity in action recognition, where static frames are sufficient to recognize an entire action video.

Shortcut opportunities have been identified in image classification and action recognition. In this thesis, we investigate whether shortcut opportunities occurs also in *long-term* action recognition datasets. Since long-term actions usually last several minutes and contain multiple short-term actions, it is commonly believed that models capable of long-term reasoning are necessary for accurate classification. However, if short-term shortcut opportunities exist, it is possible to recognize these videos solely exploiting short-term information. For example, detecting a coffee mug in a short video clip might be sufficient to recognize the long-term action of making coffee from the Breakfast dataset [27], a popular dataset of cooking videos. Discovering shortcut opportunities in long-term action recognition is important to understand the behavior of current models, identifying their limitation and exploring potential improvements.

1.3 ORGANIZATION OF THIS THESIS

In this thesis, we explore whether computer vision systems are typically aligned with their human-defined end goals. In particular, we investigate whether the training and evaluation metrics used in object detection systems conform with what humans consider good or bad object detections and how to make these models more in line with human preference. Subsequently, we examine whether long-term information is necessary to classify long-term action videos in current datasets, or if it is sufficient to exploit short-term shortcuts.

The remainder of this thesis is composed of the following original contributions:

Chapter 2 Object detectors are employed to assist humans in several applications, including industrial inspections and medial image analysis. In this chapter [28], we investigate whether the evaluation metrics of object detectors, in particular the IoU, are in line with human quality judgments. The IoU is low if the predicted object detections have different size and position with respect to the ground-truth detections. However, the IoU does not take into account if the predicted box is too small or too large, nor the shifting directions. We conduct fully controlled experiments where we ask humans to evaluate object detections with variable size and position, but same IoU. We found that humans perceived quality is higher for larger over smaller detections and that position matters for asymmetric objects. This is the first work to show that the IoU metric is insufficient to evaluate object detectors meant for human applications.

Chapter 3 In this chapter [29], we extend the evaluation in Chapter 3 to three real, widely used, object detectors. We find that these object detectors predict too large and too small bounding boxes equally often, and thus are not in line with human preference. We propose to scale the predicted detections and found that the up-scaled object detections are preferred by humans over the model predictions, even if they result in a very low AP. This result confirms the mismatch between human quality perception and object detectors evaluation metrics. It also suggests that the ground-truth object detections might not be in line with the human preference. Finally, we propose an asymmetric loss function that favors large object detections, without the need of re-annotating object detection datasets. We find that fine-tuning with the asymmetric loss results in object detections preferred over fixed up-scaling, probably due to the former being more sensitive to the object size.

Chapter 4 Chapters 2 and 3 focus on the utilization of object detections to assist human applications. In this chapter [30], we investigate whether object detections

can be used to improve human behavior recognition in minute-long videos. Under the hypothesis that focusing on the human subjects enhances human action recognition, we introduce a multi-region action recognition model that takes multiple spatial regions as input and adaptively chooses where to direct attention. We include an "actor-focused" region, centered around the person performing the activity, which we extract by means of an object detector. We also investigate whether an analogous attention mechanism in the temporal dimension helps recognizing human behavior. While multi-region attention significantly improves the results over the baseline, we surprisingly find that temporal attention does not help, and even deteriorates the performance. This result is counter-intuitive, as we would expect an auxiliary temporal model to enhance the performance by drawing attention on the most discriminative moments in a minute-long video.

Chapter 5 In complex, minute-long activities, like a football game, we would expect that not every minute is equally important. For example, in a football video a penalty kick is probably more informative than the halftime. However, in Chapter 4 we find that temporal attention does not enhance human action recognition. To understand this finding, in this chapter [31], we perform an in-depth analysis of three common long-term action recognition datasets. We find that the Breakfast and CrossTask datasets contain short-term actions that directly map to long-term action classes. We hypothesize that recognizing these short-term actions is sufficient to correctly infer the long-term classes, without the need of long-term modeling. We conduct two types of user studies, where we ask the participants to classify the long-term action in the dataset videos after seeing the full videos or short video segments. We find a very small difference in long-term action recognition performance from the two groups of participants. This shows that the videos from the three analyzed datasets do not need long-term information to be correctly classified. Computer vision algorithms are likely to make use of short-term shortcuts to correctly classify these videos, without encoding any long-term information. We recommend to use different datasets to study the problem of long-term action recognition.

Chapter 6 Current video understanding algorithms based on convolutional neural networks extract temporal information from videos through their temporal receptive field (RF). In this chapter [32], we investigate whether the temporal RF can overfit on specific long-term information at training time, in particular short-term action order. We propose Video BagNet, a 3D convolutional network with small temporal RF. We show that models with large temporal RF encode strict short-term action orders and fail when the orders at training and test time are different. On

the other hand, Video BagNet is less sensitive to permutations of the short-term actions. We find small temporal RFs perform better on the MultiTHUMOS dataset, confirming that long-term modeling is not necessary in current datasets.

Other publications Additional papers published during the research that are not integral to this thesis can be found in the List of Publications.

REFERENCES

- [1] OpenAI. *ChatGPT*. <https://openai.com/chatgpt>. 2023.
- [2] *Midjourney*. <https://www.midjourneyai.ai>. 2023.
- [3] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. “Pascal visual object classes challenge results”. In: *Available from www.pascal-network.org* 1.6 (2005), p. 7.
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [5] S. Yun, S. J. Oh, B. Heo, D. Han, J. Choe, and S. Chun. “Re-labeling imagenet: from single to multi-labels, from global to localized labels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2340–2350.
- [6] L. Bottou. “Large-scale machine learning with stochastic gradient descent”. In: *Proceedings of COMPSTAT’2010: 19th International Conference on Computational Statistics Paris France, August 22–27, 2010 Keynote, Invited and Contributed Papers*. Springer. 2010, pp. 177–186.
- [7] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).
- [9] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.* “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2020.
- [11] S. L. Smith, A. Brock, L. Berrada, and S. De. “ConvNets Match Vision Transformers at Scale”. In: *arXiv preprint arXiv:2310.16764* (2023).
- [12] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. “Selective Search for Object Recognition”. In: *International Journal of Computer Vision* 104.2 (2013), pp. 154–171. URL: <https://ivi.fnwi.uva.nl/isis/publications/2013/UijlingsIJCV2013>.

- [13] R. Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.
- [14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.
- [15] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft coco: Common objects in context”. In: *ECCV*. Springer, 2014.
- [16] J. Carreira and A. Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset”. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.
- [17] K. Hara, H. Kataoka, and Y. Satoh. “Learning spatio-temporal features with 3d residual networks for action recognition”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 3154–3160.
- [18] P. Byvshev, P. Mettes, and Y. Xiao. “Are 3D convolutional networks inherently biased towards appearance?” In: *Computer Vision and Image Understanding* 220 (2022), p. 103437.
- [19] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2630–2640.
- [20] N. Hussein, E. Gavves, and A. W. Smeulders. “Timeception for complex action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 254–263.
- [21] J. Zhou, K.-Y. Lin, H. Li, and W.-S. Zheng. “Graph-based high-order relation modeling for long-term action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8984–8993.
- [22] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *arXiv preprint arXiv:1811.12231* (2018).
- [23] W. Brendel and M. Bethge. “Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet”. In: *International Conference on Learning Representations*. 2018.
- [24] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.
- [25] M. Jain, J. C. van Gemert, and C. G. M. Snoek. “What do 15,000 Object Categories Tell Us About Classifying and Localizing Actions?” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.
- [26] Y. Li, Y. Li, and N. Vasconcelos. “Resound: Towards action recognition without representation bias”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 513–528.

- [27] H. Kuehne, A. Arslan, and T. Serre. “The language of actions: Recovering the syntax and semantics of goal-directed human activities”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 780–787.
- [28] O. Strafforello, V. Rajasekar, O. S. Kayhan, O. Inel, and J. van Gemert. “Humans disagree with the IoU for measuring object detector localization error”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 1261–1265.
- [29] O. Strafforello, O. S. Kayhan, O. Inel, K. Schutte, and J. C. van Gemert. “Aligning object detector bounding boxes with human preference”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2024.
- [30] L. Ballan, O. Strafforello, and K. Schutte. “Long-term Behaviour Recognition in Videos with Actor-focused Region Attention.” In: *VISIGRAPP (5: VISAPP)*. 2021, pp. 362–369.
- [31] O. Strafforello, K. Schutte, and J. van Gemert. “Are current long-term video understanding datasets long-term?” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2023.
- [32] O. Strafforello, X. Liu, K. Schutte, and J. van Gemert. “Video BagNet: short temporal receptive fields increase robustness in long-term action recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2023.

2

HUMANS DISAGREE WITH THE IOU FOR MEASURING OBJECT DETECTOR LOCALIZATION ERROR

The localization quality of automatic object detectors is typically evaluated by the Intersection over Union (IoU) score. In this work, we show that humans have a different view on localization quality. To evaluate this, we conduct a survey with more than 70 participants. Results show that for localization errors with the exact same IoU score, humans might not consider that these errors are equal, and express a preference. Our work is the first to evaluate IoU with humans and makes it clear that relying on IoU scores alone to evaluate localization errors might not be sufficient.

2.1 INTRODUCTION

The main difference between image classification and object detection is that an object detector also has to predict the object's location, typically indicated by a bounding box around the object. Object location can be used as a first step for a downstream task, e.g., instance segmentation [1], or human pose estimation [2]. Alternatively, in this paper, we focus on the setting where an object detection is presented to humans as an end result, where examples include visual inspection [3],

This chapter has been published as:

O. Strafforello, V. Rajasekar, O. S. Kayhan, O. Inel and J. C. van Gemert. "Humans disagree with the IoU for measuring object detector localization error". In: *IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 1261-1265

Code available at:

https://github.com/ombretta/humans_vs_IoU

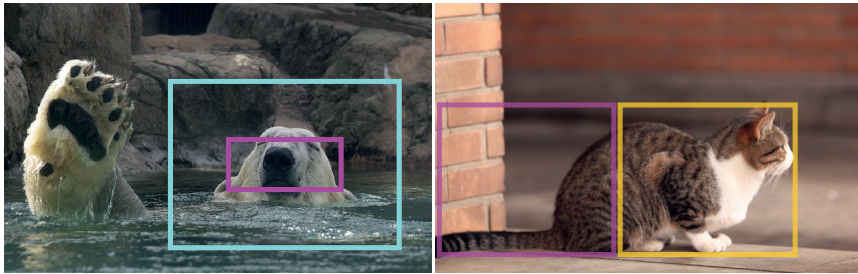


Figure 2.1: Left: Two localizations where the magenta box (0.5 IoU) is accepted, and the cyan box (0.3 IoU) is rejected by object detectors. Right: Two equally accepted localizations (0.5 IoU) by object detectors. Which boxes do you accept?

or focusing attention in medical images [4]. We do not evaluate the object detector itself [5]. Instead, we evaluate if the predicted object location by object detectors aligns with what humans consider a detected object.

Evaluating object detectors. Object detectors are commonly evaluated [5–9] with mean average precision (mAP): the mean of the per-class average precision scores. Average precision is the area under the precision-recall curve, created by ranking all detections by confidence and then checking if they are correct according to the ground truth. The detection is correct if (1) the assigned class label is correct and (2) the detection location has sufficient overlap with the ground truth. The Intersection over Union (IoU) score is used to determine the overlap. The location of a detection is correct if the IoU score is higher than a threshold, typically 0.5 or higher [6, 10]. In this paper, to the best of our knowledge, we are the first to investigate how well the IoU measure aligns with human localization quality judgments.

Human annotation for object detection. Extensive crowdsourcing studies are performed to draw bounding boxes around objects in images [11, 12] or the precise shape of the object [13, 14]. Experiments in which crowd workers validate object detections showed that annotators tend to be lenient when validating bounding boxes, *i.e.*, bounding boxes with $\text{IoU} < 0.5$ are still accepted [15]. Furthermore, analyses performed in [16] suggest that to efficiently and accurately localize all objects in an image, several crowdsourcing tasks are needed, such as verifying box correctness, verifying object presence, or naming the object. In this paper, we extend the work in [16–18] with four user studies investigating which bounding boxes humans accept and prefer.

Contributions. We make the following contributions: (1) We design four user

studies to explore what kind of detections humans prefer and accept as good detections.¹ (2) We investigate the relationship between a too small bounding box and a too large bounding box, where they both have the same IoU score. (3) We analyze the impact of object symmetry and bounding box position in human preference and acceptance of detectors' output. (4) We experiment with various object sizes (small, medium, large) and recommend future studies.

Our results show that humans disagree with IoU for measuring localization errors.

2.2 EXPERIMENTAL APPROACH

We perform four controlled experiments to evaluate the relation between IoU and human localization quality judgments and study which object detections are accepted or preferred by humans. We do not train or test any object detection models since they are highly influenced by many design choices, e.g., model parameters, dataset. Thus, our boxes are generated according to the ground truth. We relate our findings to machine-evaluated detections. For machine-evaluated detections, we use the common IoU, measuring the localization performance of the predicted box B_p with the ground truth box B_{gt} , as $\text{IoU} = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}}$.

We address two important features of object localization: (i) *Box Size* and (ii) *Box Position*, which are affected by the IoU score, in four online user studies (two studies per feature).² We also experiment with various object sizes (small - S, medium - M, large - L)³ and IoU values (0.3, 0.5, 0.7, 0.9) to study differences and similarities between humans and detection algorithms.

Procedure and participants. All studies follow the same procedure. Participants are given an example to introduce the task. The task consists of a masked image to indicate which object is investigated, the question that directly specifies the object name, and the possible answers. The images are chosen from the MS COCO dataset [10]. We ran the studies using Qualtrics⁴. The user studies have been distributed among research group members and authors' peers.

Box Size. As illustrated in Fig. 2.2, we use two different box sizes, *small* and *large*, with the same IoU score. The box aspect ratio and position is taken from the ground truth box. In the *Size Preference* study, we investigate the box size, and ask participants which box size they prefer for a detection. They can choose one

¹Data and analysis is available at https://github.com/ombretta/humans_vs_IoU.

²Ethical approval was not required - we do not collect personal identifiers.

³We adopt the definition of object size provided with the MS COCO dataset (<https://cocodataset.org/#detection-eval>).

⁴<https://www.qualtrics.com/>

option among: large box, small box or “the size of the box does not matter”. In the *Size Acceptance* study, we show either a small or a large box and ask participants if they accept or reject it as an object detection. For both studies we evaluate IoU values (0.3, 0.5, 0.7, 0.9) and include all object sizes (S, M, L). In the *Size Preference* study, we annotate 72 images, with six images per each combination between object size and IoU value. In the *Size Acceptance* study, we annotate 96 images (eight per combination).

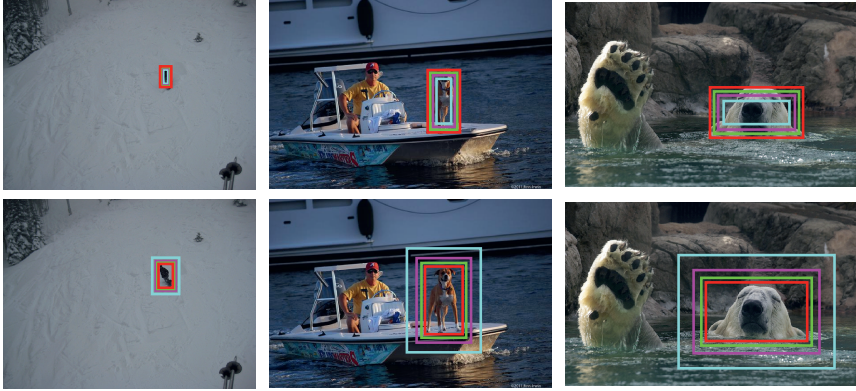


Figure 2.2: Size preference experiment. The columns indicate Small, Medium and Large object categories. The colors represent IoU scores of each box: Red (0.9), Green (0.7), Magenta (0.5) and Cyan (0.3). Top row: small bounding boxes; Bottom row: large bounding boxes. The small and large boxes of same color have the same IoU scores.

Box Position. As illustrated in Figure 2.3, we applied two positional shifts to the ground truth box, for symmetrical and asymmetrical objects, using a fixed IoU value of 0.5. Unlike the size experiment, the predicted box size is fixed and only the position of the box changes to evaluate the effect of the position. Depending on the orientation of the object, the predicted box is shifted horizontally (back, front) or vertically (top, bottom). Since symmetrical objects do not have front and back sides, we consider front as the right side and back as the left side of the object. Similarly to the size surveys, in the *Position Preference* study, we ask participants if they prefer a particular part or side of the object for detection. The *Position Acceptance* study investigates if users would accept the bounding box as a correct detection. In both position surveys, we use 20 images, which are equally distributed across object types (symmetrical, asymmetrical) and box positions (front/top, back/bottom), with 5 images per category.

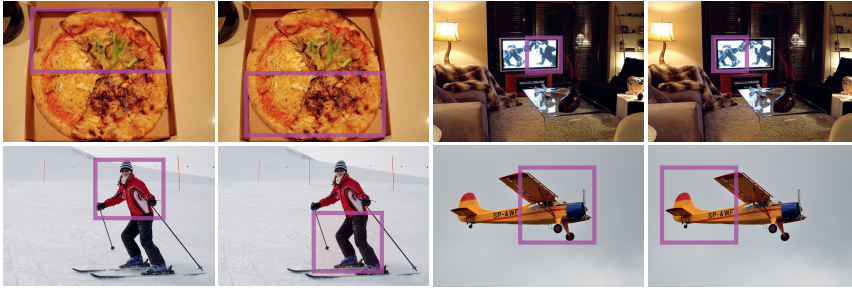


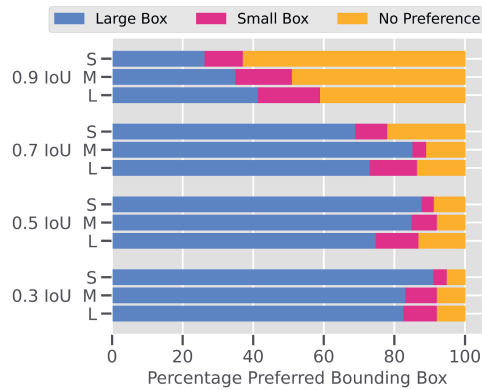
Figure 2.3: Position preference experiment. The experiments show the bounding box locations for IoU score 0.5 by shifting them horizontally or vertically. Top row: symmetrical objects; Bottom row: asymmetrical objects.

2.3 RESULTS

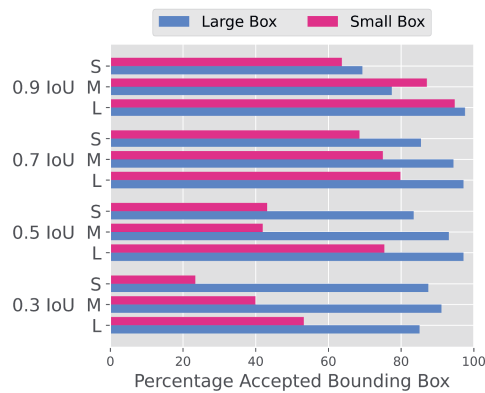
Analytical method. To study the human preference and acceptance of bounding box sizes and positions, we apply several statistical tests. We apply the Chi-square test [19] to find out if there are any associations between variables such as object size and preferred box size or IoU value and preferred box size. To understand whether differences in preference proportions (e.g., small boxes, large boxes, no preference), or acceptance proportions (e.g., front box, back box) are statistically significant, we apply the Z-test [20] and the Cochran's Q test [21]. While the Z-test can only be applied to compare two proportions, the Cochran's Q test can be applied on any number of proportions. In case of statistically significant differences, we apply a posthoc Dunn test with Bonferroni correction [22] to see which proportions are different. Since for each study we perform multiple comparisons and statistical tests, we use a lower significance threshold than 0.05 (by applying a Bonferroni correction), i.e., $\alpha = \frac{0.05}{\#tests}$.

Size Preference. Figure 2.4(a) shows, per IoU and object size, the percentage of preferred bounding box sizes. For 0.9 IoU value, people have no size preference — for each object size, the option *no preference* is either the most chosen, or similarly chosen as *large boxes*. For IoU values of 0.9, posthoc Dunn tests with Bonferroni correction show that *no preference* is statistically preferred for small and medium objects, but not for large objects. The prevalence of *no preference* is sensible: for $\text{IoU} > 0.9$, the difference in appearance between small and large boxes is subtle to the human eye.

For all other evaluated IoU values, 0.7, 0.5, 0.3, and for all three evaluated object sizes, the Cochran's Q test shows that there are statistically significant differ-



(a) Size Preference Study

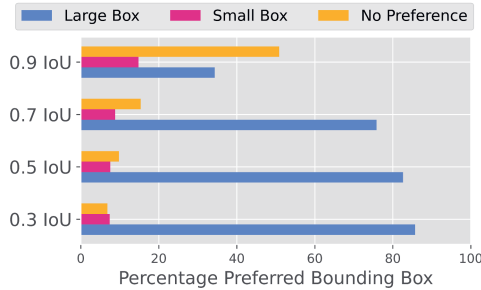


(b) Size Acceptance Study

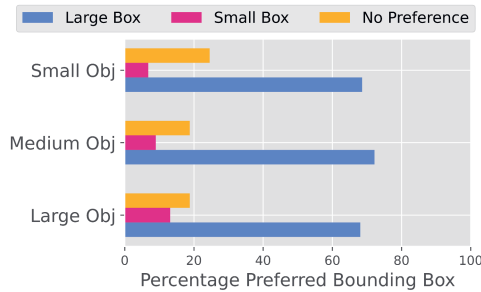
Figure 2.4: Results from studies *Size Preference* and *Size Acceptance*. a) Percentage of preferred bounding box size (small, large, no preference) for each IoU (0.3, 0.5, 0.7, 0.9) and object size (S, M, L). b) Percentage of accepted bounding box size (small, large) for each IoU and object size. The large boxes are mostly preferred and accepted by humans.

ences in the preference of boxes. Posthoc Dunn tests with Bonferroni correction indicate that *large boxes* are statistically significantly more preferred by humans. Small bounding boxes are always the least preferred while large bounding boxes are always the most preferred, irrespective of object size. We observe a gradual preference increase of *small* bounding boxes as the IoU value increases, and a

comparatively higher increase in having *no preference* (see Figure 2.5(a)). Using a Chi-square test, we found an association between the IoU value and the preferred bounding box size ($\chi^2(2)=1227.84$, $p < 0.006$). We also notice a gradual decrease in the preference of small bounding boxes with the decrease of the object size. These results are shown in Figure 2.5(b). Using a Chi-square test, we found a statistically significant association between the object size and the size of the preferred bounding box ($\chi^2(2)=62.05$, $p < 0.006$).



(a) IoU Value vs. Bounding Box Size



(b) Object Size vs. Bounding Box Size

Figure 2.5: Results from *Size Preference* study. a) Percentage of preferred bounding box size (small, large, no preference) for each IoU value (0.3, 0.5, 0.7, 0.9). b) Percentage of preferred bounding box size for each object size (S, M, L).

Size Acceptance. In Figure 2.4(b), we show the percentage of accepted *small* and *large* boxes, for each IoU value and image size. For each IoU value, the acceptance of *small* bounding boxes decreases with the decrease of object size, the smaller the object, the less accepted the *small* bounding boxes. *Large* bounding boxes are always more accepted than *small* bounding boxes, disregarding IoU values and object sizes. The exception are medium objects with 0.9 IoU, where *small* boxes

are statistically significantly more accepted ($z=-2.82$, $p < 0.008$). For the rest of the cases, *large* bounding boxes are statistically significantly more accepted than *small* bounding boxes for IoU values of 0.3, 0.5 and 0.7 and all object sizes ($p < 0.008$), but are not more accepted for neither small nor large objects with 0.9 IoU. We also found, c.f. Z-test, that (1) *large* bounding boxes are always statistically significantly accepted ($p < 0.008$) and (2) *small* bounding boxes are only statistically significantly more accepted for 0.9 and 0.7 IoU (all object sizes) and large objects with 0.5 IoU.

Position Preference. Figure 2.6(a) presents the results of the *Position Preference* user study. For symmetrical objects, participants have no preference regarding the position (*front/top* or *back/bottom*) of the bounding box, *no preference* being chosen the most. According to the Cochran's Q test, we also find that there are statistically significant differences in proportions among the three options chosen by study participants ($\chi^2(2)=268.76$, $p < 0.017$). A pairwise posthoc Dunn test with Bonferroni correction indicates that there are statistically significant differences between the proportions in which *no preference* and *front* bounding boxes are preferred ($p < 0.017$), as well as between the proportions of *no preference* and *back* bounding boxes ($p < 0.017$).

For asymmetrical objects, however, the most preferred bounding box is positioned at the *front* of the object. The Cochran's Q test shows that the difference in proportions among the three options is statistically significant ($\chi^2(2) = 576.74$, $p < 0.017$). Posthoc analysis using the Dunn test with Bonferroni correction shows that these differences are statistically significant between each two possible answers (*front* and *no preference*, *front* and *back*).

Position Acceptance. Figure 2.6(b) presents the results of the *Accepted Box Position* study. For both symmetrical and asymmetrical objects, the *front* bounding box is accepted in higher proportions than the *back* bounding box. For symmetrical objects, we found sufficient evidence, c.f. Z-test, that the proportion of *back* ($z = -7.16$, $p < 0.008$) and *front* ($z = -12.62$, $p < 0.008$) bounding boxes of being accepted is higher than the proportion of not being accepted. For asymmetrical objects, however, only *front* bounding boxes are statistically significant accepted ($z = -20.18$, $p < 0.008$). Similarly, for each object type, we analyze whether one type of bounding boxes is more accepted than the other. For both symmetrical and asymmetrical objects, the *front* bounding boxes are statistically significant more accepted than *back* bounding boxes.

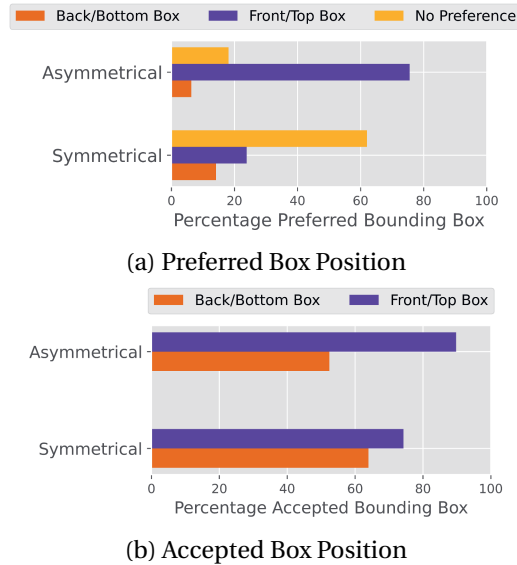


Figure 2.6: Results from studies *Position Preference* and *Position Acceptance*. a) Percentage of preferred bounding box position (front, back, no preference) for symmetrical and asymmetrical objects. b) Percentage of accepted bounding box position (front, back) for symmetrical and asymmetrical objects.

2.4 DISCUSSION

In this paper, we performed four user studies to understand which object detections are preferred and accepted by humans. We addressed two main features of object localization, namely the scale (large, small) and the position (front/top, back/bottom) of the bounding boxes, and we experimented with objects of various sizes (small, medium, large) and symmetries (symmetrical and asymmetrical).

Our studies show a statistically significant relationship between the IoU value and the preferred bounding box size, as well as between the object size and the preferred bounding box size.

Large bounding boxes are both the most preferred and the most accepted, while object detectors accept and prefer large and small boxes similarly if the boxes have the same IoU scores. We also found that for asymmetrical objects, the position of the bounding box matters for study participants, since they tend to choose bounding boxes that define or help them identify the object. This observation contrasts

current state-of-the-art object localization models [23–29], where all bounding box positions are considered correct, regardless of their orientation, when the IoU is higher than the threshold.

Object detection models, when intended for humans, should be developed in a user-centric manner *i.e.*, they should incorporate end-users preferences and comply with end-users needs. Thus, future studies should focus more on understanding which aspects of the objects should be captured by bounding boxes. The current study can also be extended by considering multiple datasets, occluded or truncated objects or images with multiple objects, as well as bounding boxes that are not centered, or which are shifted in random positions. Nevertheless, future studies should consider improving object detectors based on human preferences.

REFERENCES

- [1] K. He, G. Gkioxari, P. Dollár, and R. Girshick. “Mask r-cnn”. In: *ICCV*. IEEE/CVF. 2017.
- [2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2019).
- [3] O. S. Kayhan, B. Vredebregt, and J. C. van Gemert. “Hallucination In Object Detection—A Study In Visual Part Verification”. In: *ICIP*. IEEE. 2021.
- [4] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng. “CLU-CNNs: Object detection for medical images”. In: *Neurocomputing* 350 (2019), pp. 53–59.
- [5] D. Hoiem, Y. Chodpathumwan, and Q. Dai. “Diagnosing error in object detectors”. In: *ECCV*. Springer, 2012.
- [6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [7] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, T. Duerig, and V. Ferrari. “The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale”. In: *IJCV* (2020).
- [8] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. USA: McGraw-Hill, Inc., 1986. ISBN: 0070544840.
- [9] O. S. Kayhan and J. C. van Gemert. *Evaluating Context for Deep Object Detectors*. 2022. DOI: 10.48550/ARXIV.2205.02887. URL: <https://arxiv.org/abs/2205.02887>.
- [10] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft coco: Common objects in context”. In: *ECCV*. Springer, 2014.
- [11] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes. “Do We Need More Training Data or Better Models for Object Detection?.” In: *BMVC*. Vol. 3. 5. Citeseer. 2012.
- [12] S. Song, L. Zhang, and J. Xiao. “Robot in a room: Toward perfect object recognition in closed environments”. In: *CoRR*, *abs/1507.02703* (2015).
- [13] J. Yuen, B. Russell, C. Liu, and A. Torralba. “Labelme video: Building a video database with human annotations”. In: *CVPR*. IEEE/CVF. 2009.
- [14] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. “LabelMe: a database and web-based tool for image annotation”. In: *IJCV* (2008).

- [15] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. “We don’t need no bounding-boxes: Training object class detectors using only human verification”. In: *CVPR. IEEE/CVF*, 2016.
- [16] O. Russakovsky, L.-J. Li, and L. Fei-Fei. “Best of both worlds: human-machine collaboration for object annotation”. In: *CVPR. IEEE/CVF*. 2015.
- [17] H. Su, J. Deng, and L. Fei-Fei. “Crowdsourcing annotations for visual object detection”. In: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.
- [18] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin. “Libra r-cnn: Towards balanced learning for object detection”. In: *CVPR. IEEE/CVF*, 2019.
- [19] M. L. McHugh. “The chi-square test of independence”. In: *Biochemia medica* 23.2 (2013).
- [20] R. Schumacker. “Z test for differences in proportions”. In: *Learning statistics using R. SAGE Publications* (2017).
- [21] W. G. Cochran. “The comparison of percentages in matched samples”. In: *Biometrika* 37.3/4 (1950).
- [22] E. W. Weisstein. “Bonferroni correction”. In: <https://mathworld.wolfram.com/> (2004).
- [23] R. Girshick, J. Donahue, T. Darrell, and J. Malik. “Region-based convolutional networks for accurate object detection and segmentation”. In: *PAMI* (2015).
- [24] R. Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.
- [25] S. Ren, K. He, R. Girshick, and J. Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*. 2015.
- [26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. In: *CVPR. IEEE/CVF. IEEE/CVF*, 2016.
- [27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. “SSD: Single Shot MultiBox Detector”. In: *ECCV*. Vol. 9905. Springer, Oct. 2016, pp. 21–37. ISBN: 978-3-319-46447-3. DOI: 10.1007/978-3-319-46448-0_2.
- [28] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. “Focal loss for dense object detection”. In: *ICCV. IEEE/CVF*. 2017.
- [29] J. Redmon and A. Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).

3

ALIGNING OBJECT DETECTOR BOUNDING BOXES WITH HUMAN PREFERENCE

Previous work shows that humans tend to prefer large bounding boxes over small bounding boxes with the same IoU. However, we show here that commonly used object detectors predict large and small boxes equally often. In this work, we investigate how to align automatically detected object boxes with human preference and study whether this improves human quality perception. We evaluate the performance of three commonly used object detectors through a user study ($N = 123$). We find that humans prefer object detections that are upscaled with factors of 1.5 or 2, even if the corresponding AP is close to 0. Motivated by this result, we propose an asymmetric bounding box regression loss that encourages large over small predicted bounding boxes. Our evaluation study shows that object detectors fine-tuned with the asymmetric loss are better aligned with human preference and are preferred over fixed scaling factors. A qualitative evaluation shows that human preference might be influenced by some object characteristics, like object shape.

Object detectors identify and localize objects in an image. We focus on the common setting where detections are presented to a human by drawing a bounding box around the objects. In this paper, we evaluate how to best present object detections to humans, which is paramount for all applications that rely on showing detection to humans, such as visual inspection [1–3], anomaly detection [4–6],

This chapter has been published as:

O. Strafforello, O. S. Kayhan, O. Inel, K. Schutte and J. C. van Gemert. "Aligning object detector bounding boxes with human preference". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2024

Code available at:

https://github.com/ombretta/humans_vs_IoU

or medical imaging [7, 8]. Previous work showed that humans prefer larger over smaller boxes with the same localization error [9]. This was concluded in an on-line study with a fully controlled setup, where ground truth bounding boxes are precisely matched to the localization error. However, it is not directly clear if this controlled setting translates to the real world, where object detector outputs are imperfect. In this work, we extend [9] to real-world settings and real object detectors, which is important for reproducibility, and realistic, practical applications of scientific results.

3

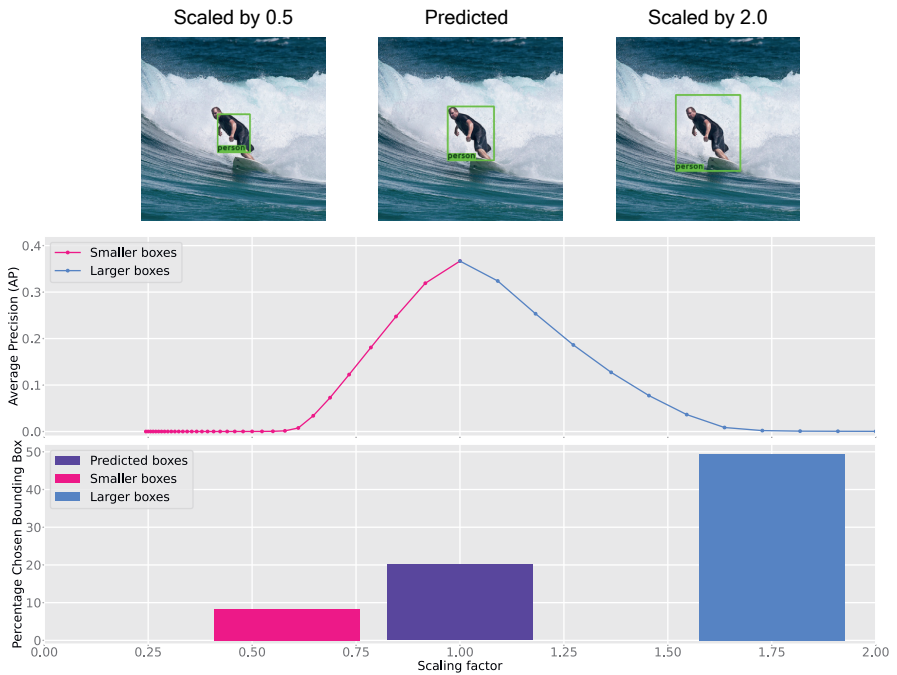


Figure 3.1: Scaling the predicted bounding box of Faster R-CNN [10] on the COCO [11] validation set. Average Precision (AP) (*top*) versus human preference (*bottom*). A scaling factor of 1.0 corresponds to the original bounding box size. Upscaling and downscaling the size of the bounding boxes severely deteriorates AP. However, our study shows that humans prefer larger bounding boxes, even if they give nearly 0 AP.

Object detectors, such as two-stage [10, 12], single stage [13–15], anchorless [16–18], and transformers-based detectors [19–22] minimize a classification loss and

a localization loss for bounding box fitting. The localization loss is symmetric for errors in bounding box size: a predicted box that is 10% too large will give the same loss as a box that is 10% too small. Here, we investigate how this symmetry affects human perception of object detections.

Object detectors are typically evaluated using average precision (AP) [23–25], which depends on the accuracy of the object classification and of the bounding box localization, as measured by the Intersection over Union (IoU) with the ground truth box. We are not the first to reconsider object detection evaluation [26–28], yet, those works all assume that a perfect-fitting bounding box is best. In contrast, here we investigate if *a perfect-fitting bounding box may not be the best box* for presenting detections to humans.

We make the following contributions: (1) We analyze three popular object detectors and find that they predict small and large bounding boxes equally often. (2) We analyze how humans perceive the predictions of the object detectors focusing on the bounding box size. As shown in Figure 3.1, we find that humans prefer up-scaled object detections, even with corresponding AP close to 0. (3) We propose an asymmetric loss function that favors the prediction of large over small boxes. Our evaluation shows that fine-tuning with the asymmetric loss better aligns object detections with human preference. All our collected data, analyses, and code are available on GitHub¹.

3.1 RELATED WORK

3.1.1 PRESENTING OBJECT DETECTIONS TO HUMANS

We take a nuanced view on evaluating object detection by identifying two distinct use-cases. Case 1: A bounding box is used as pre-processing for a follow-up algorithm such as instance segmentation [29–31], video object detection [32–34], human pose estimation [35–37], action recognition [38], etc. Case 2: A bounding box is drawn on the image, and the full image is presented directly to a human, with relevant use-cases such as visual inspection [1–3], anomaly detection [4–6], medical imaging [7, 8], etc. We argue that these two use-cases deserve different treatment. For case 1, where the bounding box is a pre-processing step, it is difficult to consider all possible follow-up algorithms, and a tightly fitting box around the object, as evaluated using IoU, seems reasonable. For case 2, however, the bounding box is the final end result and is shown to a human being. Case 2 allows directly evaluating the end result in user studies, to understand what humans actually prefer in

¹<https://github.com/ombretta/humans-vs-detectors>

their object detection. This is the focus of our paper.

3.1.2 EVALUATING OBJECT DETECTORS

All object detectors such as two-stage models [10, 12, 39, 40], single stage approaches [13–15, 41], pointwise/anchorless methods [16–18], and transformers-based detectors [19–22] are commonly evaluated [24, 25, 42, 43] with mean average precision: the mean of the per-class average precision scores. Average precision (AP) is the area under the precision-recall curve, created by ranking all detections by confidence, and then checking if a detection is correct according to the ground truth. The correctness of a detection depends on the classification: if the assigned class label is wrong, the detection is wrong. A second criterion for correctness is that the location and size of the detection have sufficient overlap with the ground truth box. For determining the overlap, the Intersection over Union (IoU) score $\frac{B_p \cap B_{gt}}{B_p \cup B_{gt}}$ is used, where B_p is the predicted bounding box, and B_{gt} is the ground truth bounding box. The location of a detection is correct if the IoU score is higher than a certain threshold, typically 0.5 or higher [11, 42]. Usually, the reported AP corresponds to a specific IoU threshold, such as 0.50 (AP50), or the average across several IoU thresholds, such as AP@[0.5 : 0.95].

We are not the first to consider object detection evaluation [26–28, 44–46], yet, those works all assume that a predicted bounding box perfectly overlapping with the ground truth bounding box is best. In contrast, we here challenge the view that a best fitting bounding box is always best for presenting detections to humans. We base our challenge on the work of Strafforello *et al.* [9] who show in precisely controlled experiments on ground truth boxes that humans prefer larger boxes over smaller boxes. In this paper, we investigate the practical ramifications of Strafforello *et al.* [9] by aligning real-world object detectors with human preference.

3.1.3 OPTIMIZING OBJECT DETECTORS

Object detectors are typically optimized using an object classification loss and a bounding box regression loss for accurate localization, by aligning the IoU of the predicted box with the ground truth box. The regression loss, usually an L2 [47] or smoothed L1 [10, 14] function, forces the box coordinates to be as close as possible to the ground truth, where the IoU is often optimized as an additional loss term [10]. Previous work proposed novel object detector losses to improve the accuracy, measured in AP. Examples include using the Absolute size IoU (AIoU) [48] and the SCALoss [49]. Other work designed a new loss term to achieve computational efficiency [50]. In our paper, we propose a simple asymmetric regression

loss function that enhances the performance of object detectors with respect to human quality judgments. Previous work used asymmetric loss in Bayesian estimation [51] and for classification [52]. To the best of our knowledge, we are the first to use an asymmetric loss for bounding box regression.

3.1.4 HUMAN ANNOTATIONS FOR OBJECT DETECTION

3

The adoption of crowdsourcing platforms such as Amazon Mechanical Turk [53] or Prolific [54] facilitated the collection of large training and testing datasets for computer vision tasks [55–60], in contrast to using in-house annotators [23, 61]. For object detection, crowdsourcing studies are extensively used to draw bounding boxes around objects that appear in images [62, 63] and videos [64] and to draw the precise shape of the object [59, 60]. To eliminate the need for clustering or averaging several bounding boxes for the same object, in [65, 66], the authors proposed a three-step workflow, where one annotator performs one step: (1) draws a bounding box around an object; (2) validates the drawn bounding box and (3) decides whether there are still objects that need to be annotated in the image. These steps are repeated until all objects in an image are annotated with bounding boxes. Experiments in which the crowd validates object detections showed that annotators tend to be lenient when validating bounding boxes, *i.e.*, bounding boxes with $\text{IoU} < 0.5$ are still accepted [67]. Furthermore, analyses performed in [68] suggest that to efficiently and accurately localize all objects in an image, several crowdsourcing tasks are needed, such as verifying box correctness, verifying object presence, or naming the object.

3.2 DO HUMANS PREFER LARGER DETECTIONS?

Previous work shows that, for equal IoU, humans prefer too large boxes over too small boxes [9]. Here, we evaluate if this has practical consequences for real object detectors. We use three popular object detectors pretrained on MS COCO: Faster R-CNN [10], RetinaNet [13], and Cascade Mask R-CNN with ResNet-50 [69] + Feature Pyramid Network [70] backbone [12, 71] all implemented in the Detectron2 library [72].

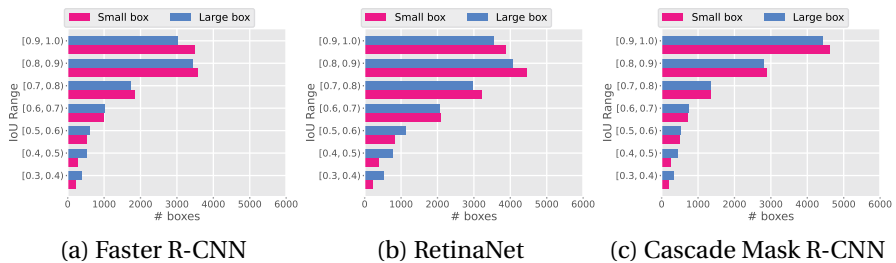


Figure 3.2: Amount of large and small bounding boxes are predicted by three object detectors on the MS COCO dataset, for seven IoU intervals, ranging from 0.3 to 1.0. For all three detectors, with higher IoU thresholds more small than large boxes are detected.

3.2.1 DO REAL DETECTORS PREDICT TOO LARGE OR TOO SMALL BOXES?

If real object detectors tend to predict too large bounding boxes, then they are already well aligned with human preference. Thus, we investigate the relative size of the predicted bounding boxes with respect to the ground truth bounding box: A *small box* has a smaller predicted area, and a *large box* has a larger predicted area. We analyze predictions on the MS COCO validation set and count the occurrences of small and large boxes.

An overview of the distribution of the predicted bounding boxes over various IoU intervals is shown in Figure 3.2. For all object detectors that we examined, there is no statistically significant difference in the number of occurrences of large and small bounding boxes. This holds for small, medium, and large objects. However, for low IoU ranges, i.e., $\text{IoU} \in [0.3, 0.6]$ for Faster R-CNN and RetinaNet and $\text{IoU} \in [0.3, 0.5]$ for Cascade Mask R-CNN, large bounding boxes are more frequent than small ones. This is due to random large bounding boxes being more likely to partially overlap with the ground truth, compared to random small boxes. Considering intermediate IoU ranges, like $\text{IoU} \in [0.6, 0.7]$, the number of occurrences of small and large boxes is not in line with the human preference found in Strafforello *et al.* [9]. That is, where humans would choose a large box over a small box with approximately 70% chance, an object detector would predict small or large with nearly equal probability.

We conclude that real object detectors generally do not predict too large boxes more often than too small boxes, and thus seem not well-aligned with human preference. In the following, we will investigate what this means for human quality

judgments of real object detectors.

3.2.2 FOR REAL OBJECT DETECTORS, DO HUMANS PREFER TOO LARGE BOXES OR TOO SMALL BOXES?

Given that, for the same IoU, humans prefer larger boxes and real object detectors do not tend to predict too large boxes, here we evaluate how humans judge re-scaled boxes. We do this through a user study, where we ask participants to evaluate five scaling factors, determined by scaling up or down the area of the predicted boxes with a factor of 1.5 and 2.0: $\{0.5, 0.67, 1.0, 1.5, 2.0\}$. Large bounding boxes are cropped to not exceed the image boundaries. Examples of bounding box scaling for a large and a small object are shown in Figure 3.3. We refer to this study as *Scaling Preference*. We ask the participants to choose the boxes they believe best identify a specific object in an image. The interface used in the user study allows the participants to select multiple options if they cannot determine a single best one. We use six random images selected from the MS COCO validation set per each combination between object size (*small*, *medium*, *large*) and IoU range. We select five IoU ranges from $0.5 \leq \text{IoU} < 0.6$ to $0.9 \leq \text{IoU} < 1.0$ that correspond to true positive predictions, for a total of 90 images. We conduct this *Scaling Preference* study on Faster R-CNN, RetinaNet, and Cascade Mask R-CNN.

Scaling the detections of a well-performing object detector results in a slight change in appearance but a significant drop in AP. For a scaling of 1, the baseline AP is 36.7%, yet a scaling of 1.5 corresponds to a $\approx 86\%$ decrease in AP. For a scaling factor of 2.0, the AP is $\approx 0\%$. Even with a more lenient IoU threshold, the AP50 decreases rapidly with both upscaling and downscaling. As shown in Table 3.1, this behavior is consistent across the three object detectors.

3.2.3 RESULTS FOR THE SCALING PREFERENCE STUDY

Table 3.2 shows the number of participants and total number of judgments for the *Scaling Preference* study. We use the Cochran's Q test [73] to determine whether there are statistically significant differences in participants' preferences regarding box sizes. In addition, we apply the posthoc Dunn tests with Bonferroni correction [74] to find what are the scaling factors that result in significant differences in users' preferences. An overview of the results is provided in Figure 3.4. We group the scaling choices into (i) "Preference for smaller boxes", if a user selected the box scaled with factor 0.67, the box scaled with factor 0.5 or both; (ii) "Preference for larger boxes", if a user selected the box scaled with factor 1.5, the box scaled with factor 2.0 or both; (iii) "Preference for original size" if a user selected only the bounding

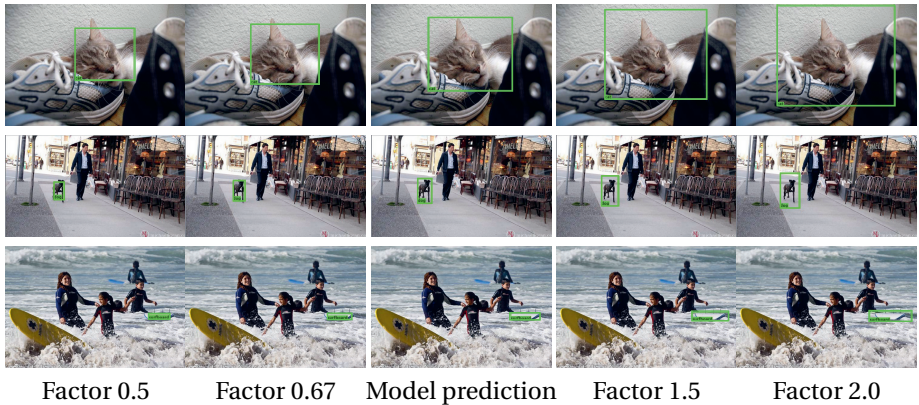


Figure 3.3: Scaling the model detections. Example of a bounding box predicted for a large object (first row), a medium object (second row) and a small object (third row) with Faster R-CNN (3rd column) and its scaled versions. In the left two images, the area of the bounding box is reduced by a scaling factor of, respectively, 0.5 and 0.67, whilst in the right two images the box area is increased by a factor of 1.5 and 2.

Scaling factor	Faster R-CNN		RetinaNet		Cascade R-CNN	
	AP	AP50	AP	AP50	AP	AP50
0.50	0.0	0.1	0.1	0.4	0.0	0.1
0.67	5.1	37.1	5.4	38.2	5.6	40.7
1.00	36.7	54.1	37.4	56.7	39.6	53.7
1.50	5.5	38.4	6.2	41.3	5.7	41.3
2.00	0.0	0.2	0.3	1.3	0.0	0.2

Table 3.1: AP (*i.e.*, $AP@[0.5 : 0.95]$) and AP50 (%) calculated for the predictions of three detectors on the MS COCO validation set and for the predicted boxes scaled with different scaling factors. Scaling the predicted boxes reduces the AP scores drastically.

box predicted by the model and (iv) "No preference" for all the remaining combinations of selections. Larger bounding boxes are consistently selected more often than small bounding boxes and than the original bounding box size for all three object detectors. This holds for different object sizes (Figure 3.4, left column), and IoU ranges (Figure 3.4, right column).

Despite the preference for larger boxes, we cannot find a statistically significant

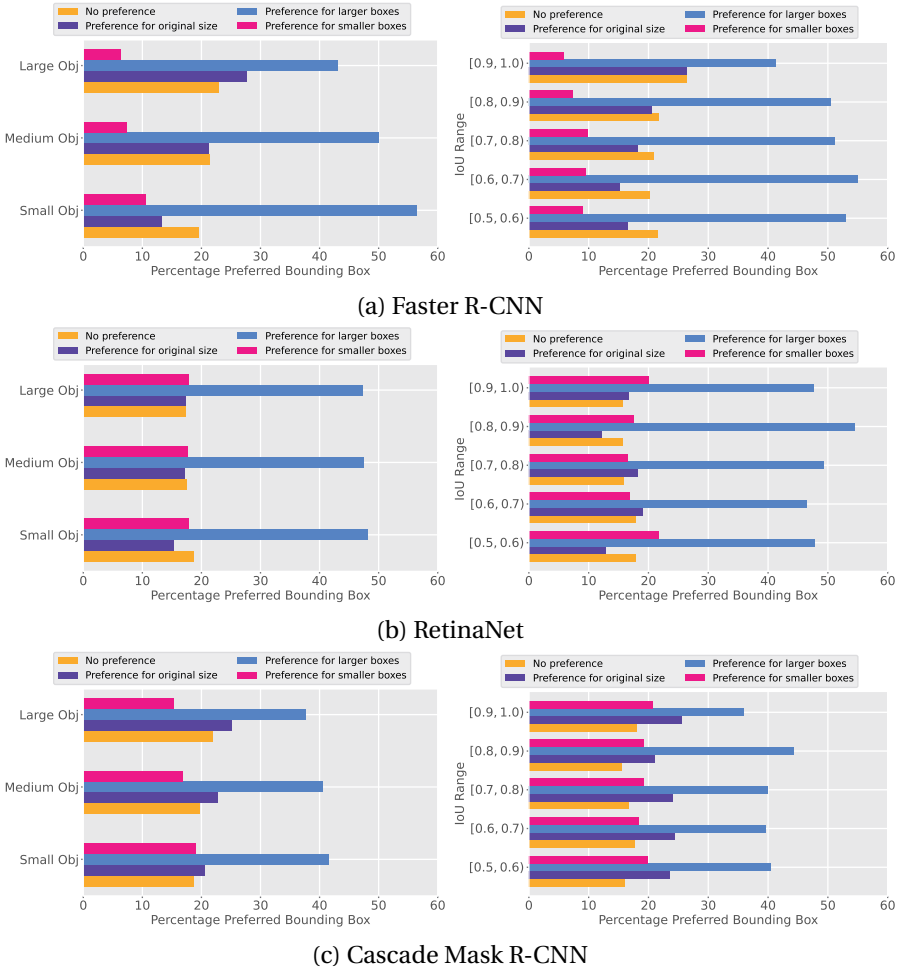


Figure 3.4: Results from the *Scaling Preference* user study. The histograms show the percentage of preferred bounding box size per object category (S, M, L) and IoU range, from $0.5 \leq \text{IoU} < 0.6$ to $0.9 \leq \text{IoU} < 1.0$, for three object detectors. The plots indicate that humans significantly prefer larger boxes.

difference between the preference for upscaling factor 1.5 and upscaling factor 2.0. For Faster R-CNN, the preference for larger boxes is composed of 56.77% of selections of both boxes scaled with factor 1.5 and 2.0; of 19.30% of selections for

scaling factor 1.5 and of 23.93% of selections for scaling factor 2.0. Here, Dunn’s test shows no statistically significant difference between the preference for scaling factor 1.5 and scaling factor 2.0. This means that larger boxes are preferred, but there is no single best upscaling factor. One exception holds for the bounding boxes for small objects predicted with Faster R-CNN: in this case, scaling factor 2.0 is preferred over scaling factor 1.5 (Dunn’s $\alpha \approx 0$). This preference is an indicator that, for small objects, scaling the bounding box with a large scaling factor, like 2.0, results in more satisfactory detections. A majority of votes for the largest box for small objects, albeit not statistically significant, is observed for the other object detectors. It is noticeable how larger bounding boxes are preferred to bounding boxes predicted with high IoU. This indicates that, for representative images of the diverse MS COCO dataset, humans are likely to prefer bounding boxes larger than the ground truth bounding boxes.

	Faster R-CNN	RetinaNet	Cascade Mask R-CNN
Participants	39	36	48
Judgments	5400	5220	5632

Table 3.2: Overview of participants and their judgments in the scaling preference study.

3.3 ASYMMETRIC REGRESSION LOSS TO ENCOURAGE LARGER DETECTIONS

We find that humans consistently prefer larger object detections, while object detectors predict large and small boxes equally often. We propose an asymmetric bounding box regression loss that encourages larger detections. Our asymmetric loss is obtained by a simple modification of the smooth L_1 localization loss function used in standard object detectors. We use the asymmetry term α to increase the loss value when the predicted area is smaller than the ground truth area and decrease the loss value when is larger. The asymmetric loss is given by

$$\text{Asymmetric } L_{1,\text{smooth}} = \begin{cases} \frac{1}{2\sqrt{\alpha}\beta}x^2, & \text{if } 0 \leq x < \beta \\ \frac{\sqrt{\alpha}}{2\beta}x^2, & \text{if } -\beta < x < 0 \\ \frac{1}{\sqrt{\alpha}}x - \frac{\beta}{2\sqrt{\alpha}}, & \text{if } x \geq \beta \\ -\sqrt{\alpha}x - \frac{\sqrt{\alpha}\beta}{2}, & \text{if } x \leq -\beta \end{cases} \quad (3.1)$$

The α represents the asymmetry term, β determines the standard smoothing interval in which the L_1 loss becomes quadratic, and x is the input to the loss function, which is the difference between the predicted height/width and the ground truth values, $x = x_{\text{pred}} - x_{\text{GT}}$. As shown in Figure 3.5, the asymmetric loss is identical to the smooth L_1 loss when $\alpha = 1$. We use the asymmetric loss function for the regression of the boxes' height and width.

We fine-tune Faster R-CNN, RetinaNet, and Cascade R-CNN on MS COCO for 100k iterations with the asymmetric loss. As a result, the fine-tuned models are more likely to predict larger boxes over smaller boxes. Figure 3.6 shows the percentage of large detections for different α values. Similarly to the fixed scaling factors in the *Scaling Preference* study in section 3.2, we observe a decrease in the AP with the increase of large detections. With $\alpha = 10$, we obtain 80% to 90% large predictions without compromising AP too much.

We measure the average size increase of the predicted bounding boxes compared to the ground truth. As shown in Figure 3.7, increasing the α coefficient results in an increase of the average box size, for all three models and object sizes. The models fine-tuned with $\alpha = 10$ return detections scaled compared to ground truth, on average by factors 1.21 ± 0.24 for Faster R-CNN, 1.21 ± 0.25 for RetinaNet, and 1.19 ± 0.22 for Cascade R-CNN, while fine-tuning with $\alpha = 100$ results in average scaling of 1.41 ± 0.26 for Faster R-CNN, 1.34 ± 0.28 for RetinaNet, and 1.39 ± 0.24 for Cascade R-CNN. It is noticeable that the size of small objects' detections increases more than for medium and large objects. This is mostly due to small bounding boxes having more opportunity for expansion in the image, while large objects' boxes are already close to the image boundaries.

3.3.1 DOES THE ASYMMETRIC L_1 LOSS LEAD TO DETECTIONS CLOSER TO THE HUMAN PREFERENCE?

We conduct a final user study to investigate whether adopting the asymmetric L_1 loss results in detections closer to human preference. In the evaluation study, we include the detections from the original pretrained Faster R-CNN ($\alpha = 1$), the detections from Faster R-CNN fine-tuned with $\alpha = 10$ and 100 , and the detections scaled by a fixed factor 1.5 , which was one of the preferred options in the *Scaling Preference* study 3.2. These values for α are chosen to have detection sizes that notably differ from the Faster R-CNN baseline (Figure 3.7). We ask users to compare the four different detections for the same object and choose the one that, in their opinion, best identifies the object. We include 45 detections, equally sampled from the three object categories (*small*, *medium*, *large*). We conduct the study on Amazon Mechanical Turk [53] and collect 660 judgments.

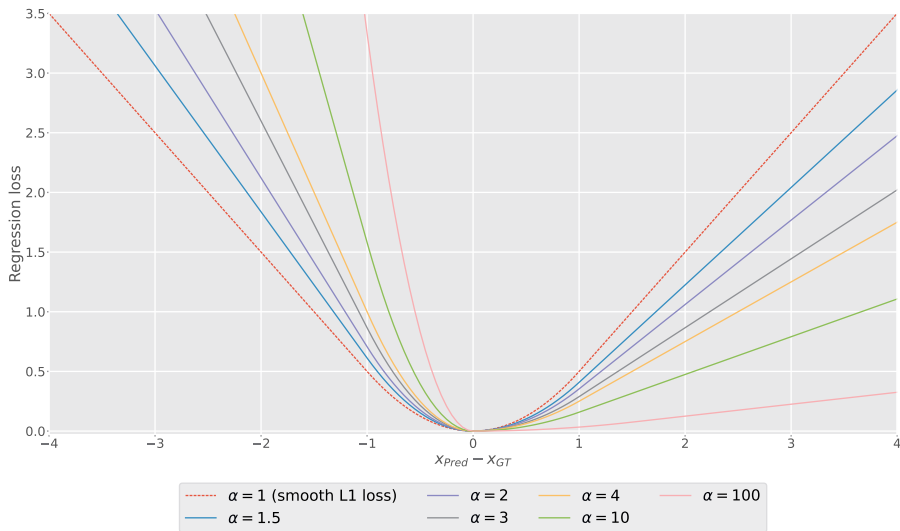


Figure 3.5: Asymmetric smooth L_1 loss with different α . Larger bounding boxes are penalized less than smaller predictions.

The results are summarized in Table 3.3. The Cochran's Q test reveals statistically significant differences between the proportions of preferred object detections. The detections obtained by fine-tuning with asymmetric loss, $\alpha = 10$, are always the most preferred. This preference is statistically significant when considering all object categories and small objects (Dunn's $\alpha \leq 0.001$). In the other cases, fine-tuning with $\alpha = 10$ is significantly more preferred than scaling with a fixed factor (Dunn's $\alpha \approx 0$), thus confirming the advantage of the asymmetric loss over fixed scaling.

The preference for the asymmetric loss over fixed scaling is likely due to the fixed scaling factor upscaling all boxes equally, irrespective of the object size. Conversely, using the asymmetric loss results in boxes upscaled more for small objects than for medium and large objects, as illustrated in Figure 3.7. This might lead to higher human preference, since large objects are already easily identifiable with a tighter box. In fact, as shown in Table 3.3, the fixed scaling 1.5 is almost never chosen for large objects. In addition, we observe that the most preferred option — asymmetric loss with $\alpha = 10$ — leads to detections that are, on average, larger than the ground truth by a factor between 1.1 and 1.5. We hypothesize that the optimal scaling factor might lie within this range. Another potential reason why scaling by 1.5 is less preferred is that the fixed scaling strategy retains the aspect ratio of the original predicted box. This aspect ratio may not be optimal when upscaling the

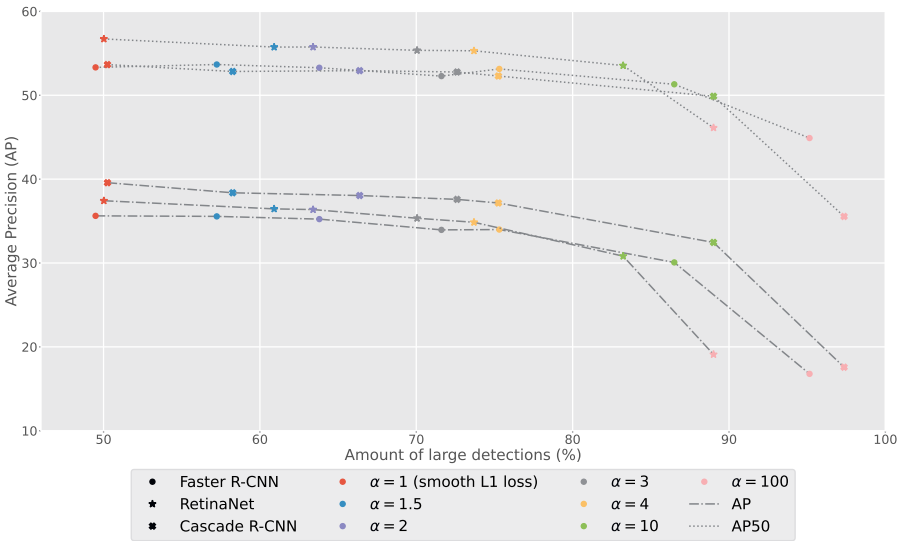


Figure 3.6: Average Precision (AP) as a function of the amount of predicted boxes that are larger than the ground truth boxes. The percentage of large detections increases with the α parameter, while the AP decreases.

boxes. In contrast, the asymmetric loss function imposes fewer constraints on the aspect ratio.

Overall, fine-tuning the models with our asymmetric L_1 loss results in detections closer to human preference. We suggest adopting this loss when object detections are meant to be presented to humans.

Object cat.	# judgments	Chosen object detection (%)			
		$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	Scal. fact. 1.5
All	660	27.4	37.1	21.2	14.2
Small	229	17.0	31.9	26.6	24.5
Medium	215	27.9	39.5	16.7	15.8
Large	216	38.0	40.3	19.9	1.9

Table 3.3: Users’ preferred object detections (%), computed with Faster R-CNN fine-tuned with the asymmetric loss function or up-scaled with factor 1.5. Fine-tuning with $\alpha = 10$ is always the most preferred option.

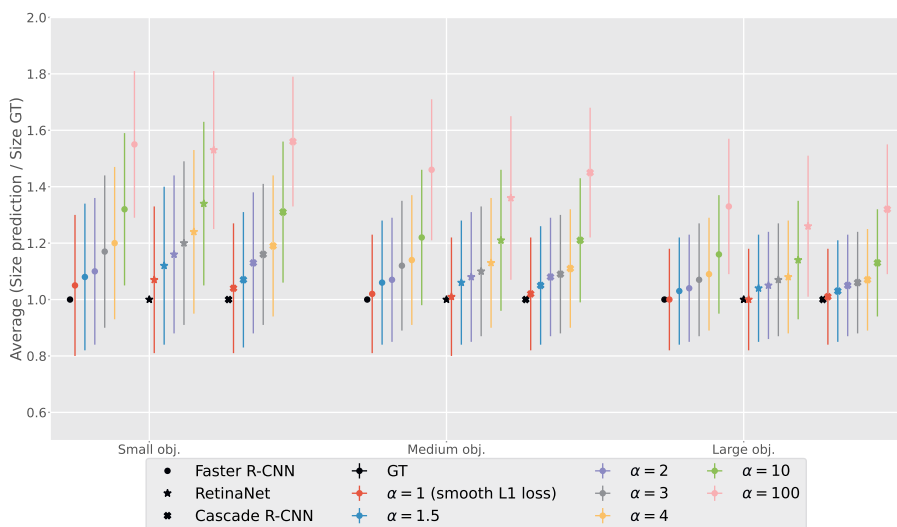


Figure 3.7: Bounding box size increases after fine-tuning object detectors with the asymmetric smooth L_1 loss with parameter α .

3.3.2 QUALITATIVE ANALYSIS OF THE PREFERRED BOXES

We manually analyze the results obtained from the user evaluation of the asymmetric loss and illustrate some representative examples in Figure 3.8. We notice that the tight bounding boxes predicted by the Faster R-CNN baseline, namely, trained with $\alpha = 1$, are generally preferred for large objects, e.g., the *cat* in the first row. Preference for $\alpha = 1$ also occurs when there are multiple objects behind or in the proximity of the object of interest. In the image on row 2 of Figure 3.8, larger bounding boxes partly include the *chair* behind the one of interest. In this case, tight boxes delineate better the subject of focus.

Slightly larger boxes, obtained by fine-tuning with our asymmetric loss, $\alpha = 10$, are preferred when small parts of the object are not contained in the tight bounding box predicted by Faster R-CNN (e.g., the *candle* on the birthday cake in Figure 3.8, row 3), or partly covered by the box contour itself, like the ears and tail of the *cat* in row 4. We hypothesize that predicted tight bounding boxes leave out possible object protrusions, despite resulting in high AP. Presumably, humans prefer large boxes because they can include the whole object. Additionally, in the presence of a uniform background (e.g., the green grass behind the cat in Figure 3.8, row 4), humans are generally less concerned if the bounding box is slightly larger.



Figure 3.8: Example of human preferences obtained from the user evaluation of the asymmetric loss. The columns show the percentage of users who prefer the bounding boxes obtained by the original pretrained Faster R-CNN ($\alpha = 1$), after fine-tuned with α 10 and 100, or scaling by a fixed factor 1.5. Generally, humans prefer tight boxes for large objects (first row) and when the object of interest overlaps with other objects (second row). Slightly larger boxes, obtained with asymmetric loss, $\alpha = 10$ or 100, are preferred when small parts of the object protrude outside too tight bounding boxes (e.g., the candle on the birthday cake, third row), or partly covered by the box line itself (fourth row). Large boxes ($\alpha = 1$ or scaling factor 1.5) are chosen for very small objects (fourth and fifth row). Finally, we found no preference when all bounding boxes are too visually similar (last row).

Similarly, the asymmetric loss makes it more likely to include all the small protruding parts of the objects in the predicted boxes.

We observe that the preference for larger boxes, obtained by scaling with factor 1.5 or with the asymmetric loss $\alpha = 100$ occurs when the objects of interest are very small, e.g., the *mouse* and the *person* walking on the street (row 5 and 6, Figure 3.8). Finally, in a few cases, the original Faster R-CNN detector, the detectors fine-tuned with the asymmetric loss or manually scaled result in very similar boxes, indistinguishable by a human eye. In this situation, we observe no clear human preference, as for the *person* in the last row in Figure 3.8.

The qualitative analysis suggests that there exists a relationship between the object characteristics, especially size (already observed in [9]) and shape, and the preferred bounding box size. We leave the investigation of the factors that determine the user preference for future work.

3.4 CONCLUSION

Prior work [9] shows that humans prefer larger boxes in a fully controlled setup. In this paper, we confirm this result in practice, with real detectors. We evaluate the bounding boxes predicted by three popular object detectors. We find that the object detectors predict large and small bounding boxes equally often, therefore are not aligned with the human preference found in [9]. In addition, humans consistently prefer larger bounding boxes over the predicted boxes, even with AP approximately zero. Therefore, we recommend being careful with AP scores when object detectors are intended for human use: a high AP does not automatically correspond to high human preference.

It is noticeable how the preference occurs even for bounding boxes predicted with high IoUs: this suggests that humans are likely to prefer larger bounding boxes compared to tight ground truth bounding boxes.

We propose an asymmetric loss function that encourages detectors to predict large boxes more often than small boxes, without having to re-annotate the training images. Our user evaluation shows that fine-tuning with the asymmetric loss results in object detections more aligned with human preference. After qualitatively analyzing the results collected from our study, we hypothesize that the human preference is affected by the object characteristics, such as shape and size. For example, generally tight boxes are preferred for large objects, while larger boxes are preferred for small objects. Further investigation into these observations may be considered in the future.

REFERENCES

- [1] O. S. Kayhan, B. Vredebregt, and J. C. van Gemert. “Hallucination In Object Detection—A Study In Visual Part Verification”. In: *ICIP*. IEEE. 2021.
- [2] Y. Li, H. Trinh, N. Haas, C. Otto, and S. Pankanti. “Rail component detection, optimization, and assessment for automatic rail track inspection”. In: *IEEE Transactions on Intelligent Transportation Systems* 15.2 (2013).
- [3] D. Mery and A. K. Katsaggelos. “A logarithmic X-ray imaging model for baggage inspection: Simulation and object detection”. In: *CVPRW*. IEEE/CVF. 2017.
- [4] A. Basharat, A. Gritai, and M. Shah. “Learning object motion patterns for anomaly detection and improved object detection”. In: *CVPR*. IEEE/CVF. 2008.
- [5] K. Doshi and Y. Yilmaz. “Fast unsupervised anomaly detection in traffic videos”. In: *CVPRW*. IEEE/CVF. 2020.
- [6] X. Li, W. Li, B. Liu, Q. Liu, and N. Yu. “Object-oriented anomaly detection in surveillance videos”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018.
- [7] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng. “CLU-CNNs: Object detection for medical images”. In: *Neurocomputing* 350 (2019), pp. 53–59.
- [8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017).
- [9] O. Strafforello, V. Rajasekar, O. S. Kayhan, O. Inel, and J. van Gemert. “Humans disagree with the IoU for measuring object detector localization error”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 1261–1265.
- [10] S. Ren, K. He, R. Girshick, and J. Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*. 2015.
- [11] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft coco: Common objects in context”. In: *ECCV*. Springer, 2014.
- [12] Z. Cai and N. Vasconcelos. “Cascade R-CNN: Delving Into High Quality Object Detection”. In: *CVPR*. IEEE/CVF. IEEE/CVF, 2018.
- [13] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. “Focal loss for dense object detection”. In: *ICCV*. IEEE/CVF. 2017.

- [14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. “SSD: Single Shot MultiBox Detector”. In: *ECCV*. Vol. 9905. Springer, Oct. 2016, pp. 21–37. ISBN: 978-3-319-46447-3. DOI: 10.1007/978-3-319-46448-0_2.
- [15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. In: *CVPR. IEEE/CVF. IEEE/CVF*, 2016.
- [16] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. “Centernet: Keypoint triplets for object detection”. In: *ICCV. IEEE/CVF*. 2019.
- [17] H. Law and J. Deng. “Cornersnet: Detecting objects as paired keypoints”. In: *ECCV*. Springer, 2018.
- [18] X. Zhou, J. Zhuo, and P. Krahenbuhl. “Bottom-up object detection by grouping extreme and center points”. In: *CVPR. IEEE/CVF*. 2019.
- [19] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk. *Toward Transformer-Based Object Detection*. 2020. arXiv: 2012.09958 [cs.CV].
- [20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. “End-to-End Object Detection with Transformers”. In: *Lecture Notes in Computer Science* (2020), pp. 213–229. ISSN: 1611-3349. DOI: 10.1007/978-3-030-58452-8_13. URL: http://dx.doi.org/10.1007/978-3-030-58452-8_13.
- [21] Z. Dai, B. Cai, Y. Lin, and J. Chen. “UP-DETR: Unsupervised Pre-Training for Object Detection With Transformers”. In: *CVPR. IEEE/CVF*. 2021, pp. 1601–1610.
- [22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. “Deformable DETR: Deformable Transformers for End-to-End Object Detection”. In: *arXiv preprint arXiv:2010.04159* (2020).
- [23] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. “The pascal visual object classes (voc) challenge”. In: *IJCV* (2010).
- [24] D. Hoiem, Y. Chodpathumwan, and Q. Dai. “Diagnosing error in object detectors”. In: *ECCV*. Springer, 2012.
- [25] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, T. Duerig, and V. Ferrari. “The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale”. In: *IJCV* (2020).
- [26] N. Chavali, H. Agrawal, A. Mahendru, and D. Batra. “Object-proposal evaluation protocol is ‘gameable’”. In: *CVPR. IEEE/CVF*, 2016.
- [27] D. Feng, Z. Wang, Y. Zhou, L. Rosenbaum, F. Timm, K. Dietmayer, M. Tomizuka, and W. Zhan. “Labels are not perfect: Inferring spatial uncertainty in object detection”. In: *IEEE Transactions on Intelligent Transportation Systems* (2021).
- [28] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. “What makes for effective detection proposals?” In: *IEEE transactions on pattern analysis and machine intelligence* 38.4 (2015).

- [29] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. “Yolact: Real-time instance segmentation”. In: *ICCV*. IEEE/CVF. 2019.
- [30] K. He, G. Gkioxari, P. Dollár, and R. Girshick. “Mask r-cnn”. In: *ICCV*. IEEE/CVF. 2017.
- [31] Y. Shen, L. Cao, Z. Chen, B. Zhang, C. Su, Y. Wu, F. Huang, and R. Ji. “Parallel Detection-and-Segmentation Learning for Weakly Supervised Instance Segmentation”. In: *ICCV*. IEEE/CVF. 2021.
- [32] Y. Chen, Y. Cao, H. Hu, and L. Wang. “Memory Enhanced Global-Local Aggregation for Video Object Detection”. In: *CVPR*. IEEE/CVF. 2020.
- [33] L. Han, P. Wang, Z. Yin, F. Wang, and H. Li. “Context and Structure Mining Network for Video Object Detection”. In: *IJCV* (2021).
- [34] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang. “New Generation Deep Learning for Video Object Detection: A Survey”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2021).
- [35] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2019).
- [36] K. Sun, B. Xiao, D. Liu, and J. Wang. “Deep High-Resolution Representation Learning for Human Pose Estimation”. In: *CVPR*. IEEE/CVF. 2019.
- [37] B. Xiao, H. Wu, and Y. Wei. “Simple baselines for human pose estimation and tracking”. In: *ECCV*. Springer, 2018.
- [38] L. Ballan, O. Strafforello, and K. Schutte. “Long-term Behaviour Recognition in Videos with Actor-focused Region Attention.” In: *VISIGRAPP (5: VISAPP)*. 2021, pp. 362–369.
- [39] R. Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.
- [40] R. Girshick, J. Donahue, T. Darrell, and J. Malik. “Region-based convolutional networks for accurate object detection and segmentation”. In: *PAMI* (2015).
- [41] J. Redmon and A. Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).
- [42] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [43] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. USA: McGraw-Hill, Inc., 1986. ISBN: 0070544840.
- [44] R. Padilla, S. L. Netto, and E. A. da Silva. “A survey on performance metrics for object-detection algorithms”. In: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE. 2020.

- [45] D. K. Prasad, H. Dong, D. Rajan, and C. Quek. "Are object detection assessment criteria ready for maritime computer vision?" In: *IEEE Transactions on Intelligent Transportation Systems* 21.12 (2019).
- [46] A. Sobti, V. Mavi, M. Balakrishnan, and C. Arora. "VmAP: A Fair Metric for Video Object Detection". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021.
- [47] J. Redmon and A. Farhadi. "YOLO9000: Better, Faster, Stronger". In: *CVPR. IEEE/CVF*, 2017, pp. 7263–7271.
- [48] D. Tian, Y. Han, S. Wang, X. Chen, and T. Guan. "Absolute size IoU loss for the bounding box regression of the object detection". In: *Neurocomputing* 500 (2022), pp. 1029–1040. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2022.06.018>. URL: <https://www.sciencedirect.com/science/article/pii/S0925231222007378>.
- [49] T. Zheng, S. Zhao, Y. Liu, Z. Liu, and D. Cai. "Scaloss: Side and corner aligned loss for bounding box regression". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 3. 2022, pp. 3535–3543.
- [50] D. Aswal, P. Shukla, and G. C. Nandi. "Designing effective power law-based loss function for faster and better bounding box regression". In: *Machine Vision and Applications* 32.4 (2021), p. 87.
- [51] A. Basu and N. Ebrahimi. "Bayesian approach to life testing and reliability estimation using asymmetric loss function". In: *Journal of statistical planning and inference* 29.1-2 (1991), pp. 21–31.
- [52] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor. "Asymmetric loss for multi-label classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 82–91.
- [53] *Amazon Mechanical Turk*. <https://www.mturk.com/>. Accessed: 2023-07-05.
- [54] *Prolific*. <https://www.prolific.co>.
- [55] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. "Activitynet: A large-scale video benchmark for human activity understanding". In: *CVPR*. 2015.
- [56] R. Di Salvo, D. Giordano, and I. Kavasidis. "A crowdsourcing approach to support video annotation". In: *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*. ACM. 2013.
- [57] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.* "Visual genome: Connecting language and vision using crowdsourced dense image annotations". In: *IJCV* (2017).
- [58] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, *et al.* "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey". In: *Monthly Notices of the Royal Astronomical Society* (2008).

- [59] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. “LabelMe: a database and web-based tool for image annotation”. In: *IJCV* (2008).
- [60] J. Yuen, B. Russell, C. Liu, and A. Torralba. “Labelme video: Building a video database with human annotations”. In: *CVPR. IEEE/CVF*. 2009.
- [61] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. “Sun database: Large-scale scene recognition from abbey to zoo”. In: *2010 IEEE computer society conference on Computer Vision and Pattern Recognition*. 2010.
- [62] S. Song, L. Zhang, and J. Xiao. “Robot in a room: Toward perfect object recognition in closed environments”. In: *CoRR, abs/1507.02703* (2015).
- [63] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes. “Do We Need More Training Data or Better Models for Object Detection?”. In: *BMVC*. Vol. 3. 5. Citeseer. 2012.
- [64] C. Vondrick, D. Patterson, and D. Ramanan. “Efficiently scaling up crowdsourced video annotation”. In: *IJCV* (2013).
- [65] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. “ImageNet Large Scale Visual Recognition Challenge”. In: *IJCV* 115 (2015), pp. 211–252.
- [66] H. Su, J. Deng, and L. Fei-Fei. “Crowdsourcing annotations for visual object detection”. In: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.
- [67] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. “We don’t need no bounding-boxes: Training object class detectors using only human verification”. In: *CVPR. IEEE/CVF*, 2016.
- [68] O. Russakovsky, L.-J. Li, and L. Fei-Fei. “Best of both worlds: human-machine collaboration for object annotation”. In: *CVPR. IEEE/CVF*. 2015.
- [69] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [70] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. “Feature Pyramid Networks for Object Detection”. In: *CVPR* (2017). DOI: 10.1109/cvpr.2017.106. URL: <http://dx.doi.org/10.1109/CVPR.2017.106>.
- [71] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. “Mask R-CNN”. In: *CoRR abs/1703.06870* (2017). arXiv: 1703.06870. URL: <http://arxiv.org/abs/1703.06870>.
- [72] Y. Wu, A. Kirillov, F. Massa, W. Lo, and R. Girshick. *Detectron2*. <https://github.com/facebookresearch/detectron2>. <https://github.com/facebookresearch/detectron2>. 2019.
- [73] W. G. Cochran. “The comparison of percentages in matched samples”. In: *Biometrika* 37.3/4 (1950).
- [74] E. W. Weisstein. “Bonferroni correction”. In: <https://mathworld.wolfram.com/> (2004).

4

LONG-TERM BEHAVIOUR RECOGNITION IN VIDEOS WITH ACTOR-FOCUSED REGION ATTENTION

Long-term activities involve humans performing complex, minutes-long actions. Differently than in traditional action recognition, complex activities are normally composed of a set of sub-actions, that can appear in different order, duration, and quantity. These aspects introduce a large intra-class variability, that can be hard to model. Our approach aims to adaptively capture and learn the importance of spatial and temporal video regions for minutes-long activity classification. Inspired by previous work on Region Attention, our architecture embeds the spatio-temporal features from multiple video regions into a compact fixed-length representation. These features are extracted with a 3D convolutional backbone specially fine-tuned. Additionally, driven by the prior assumption that the most discriminative locations in the videos are centered around the human that is carrying out the activity, we introduce an Actor Focus mechanism to enhance the feature extraction both in training and inference phase. Our experiments show that the Multi-Regional fine-tuned 3D-CNN, topped with Actor Focus and Region Attention, largely improves the performance of baseline 3D architectures, achieving state-of-the-art results on Breakfast, a well known long-term activity recognition benchmark.

This chapter has been published as:

L. Ballan, O. Strafforello and K. Schutte. “Long-term Behaviour Recognition in Videos with Actor-Focused Region Attention”. In: *VISIGRAPP (5: VISAPP)*. 2021, pp. 362-369

4.1 INTRODUCTION

Long-term activity recognition is getting increasing attention in the Computer Vision community as it allows for important applications related to video surveillance and sport video analysis. However, this task is intrinsically complex because of the long duration of the videos, the variability in the activities composition and the visual complexity of video frames from real world scenarios. Inspired by previous work on Region Attention [1], we introduce a model that can adaptively select and focus on the video regions that are most discriminative for the complex activity classification.

Our method is driven by two assumptions. Firstly, not all the locations and the moments in the videos are equally important. The activity "preparing cereal bowl", for example, has a precise location in the video frames. Other locations belong to the background, namely regions where the activity does not happen. Background locations might show "distracting" elements that might induce to misclassify the activity. Similarly, a correct classification of a cooking activity might be possible just by looking at the last seconds of the videos, that are likely to show the ready dish. On the contrary, some less informative moments might occur elsewhere, for instance when the cook is looking for the ingredients. Following this assumption, we introduce a Region Attention module, that can explicitly choose among multiple spatial and temporal input regions. This setting acts as a natural data augmentation strategy, and allows to retain only the information that is relevant for the classification.

The second assumption is that the most discriminative spatial regions in the videos are the ones placed around the actor that is accomplishing the activity. For example, for cooking activities, the ingredients and the utensils that are characteristic of the actions, are those that the cook interacts with. Therefore, focusing on the cook should give sufficient information to understand what dish is being made. Hence, we introduce an Actor Focus mechanism that allows the model to explicitly center the attention on the actor.

Due to the large intra-class variability, modelling long-term activities can be difficult. The recent solutions in the literature involve 3D-CNNs as effective spatio-temporal feature extractors [2], combined with additional modules that further process the features in the temporal dimension, including temporal convolution [3] and self-attention [4–6]. Even though these works reached competitive results in common long-term activities benchmarks, we argue that the performance of these models is heavily influenced by the quality of the 3D-CNNs backbone training. Despite their potential, 3D-CNN architectures are characterized by the downside of having a large amount of parameters that makes the learning process extremely data hungry. Since the datasets for long-term activities are limited in size

[7–9] learning general video representations with these models without overfitting on the training set is unfeasible. That is why our approach based on multiple regions is crucial to reach better generalization. We show that the combination of an optimal backbone fine-tuning, augmented with the multiple regions, with the Region Attention method and the Actor Focus mechanism achieves state-of-the-art results on the Breakfast Actions Dataset benchmark [7].

4.2 RELATED WORK

Although a wide range of solutions for short-range action recognition have been proposed [2, 10, 11], these are not necessarily transferable to long-term activity recognition, as the two data types are fundamentally different. Short actions (or *unit-actions*), such as "cutting" or "pouring" are limited in duration and consist of a single, possibly periodic, movement. Because of this, they are easily recognizable by looking at a small number of frames, sometimes even one [12]. On the contrary, long-term activities are composed by a collection of unit-actions, where some of them might be shared among different classes. For example, the action "pouring" belongs both to the classes "making tea" and "making coffee". Because of this, it is not possible to classify a complex activity by looking at a specific moment, but the whole time span should be considered. Therefore, more sophisticated architectures are required.

4.2.1 LONG-TERM MODELLING

The majority of the recently proposed works on long-term modelling enhance the exploitation of the temporal dimension. Timeception [3], for example, achieves this with multi-scale temporal convolutions which learn flexibly long-term temporal dependencies. Similarly, [13] consider different temporal extents of video representations at the cost of decreased spatial resolution. [14] propose a long-term feature bank of information extracted over the entire span of videos as context information in support to 3D-CNNs. [15] rely on STIP (Spatio-Temporal Interest Points) features weighted by their spatio-temporal probability. Another example of temporal reasoning is provided by the TRN (Temporal Relation Network) [16], that learns dependencies between video frames, at both short-term and long-term timescales. Conditional Gating adopted in TimeGate [17] enables a differentiable sampling of video segments, to discard redundant information and achieve computational efficiency. According to another recent thread, supported in VideoGraph [4] and [18], a thorough representation of complex activities can be achieved

by explicitly modelling the human-object and object-object interactions across time. The VideoGraph method learns this type of information through a fixed set of latent concepts depicting the activity evolution, whereas [19] address directly the object-object interactions, embedding them in a graph structure.

Among the most performing work that utilizes the Breakfast dataset, [5] propose a new kind of convolutional operation which is invariant to the temporal permutations within a local window. Their proposed model is better suited to handling the weak temporal structure and variable order of the unit-actions that compose the long-term activities. On the other hand, ActionVlad [20] develops a system that pools jointly across spatio-temporal features provided by a two-stream network. Finally, Non-local Nets [6] provide a building block for many deep architectures: computing the response at a position as a weighted sum of the features at all positions, they capture long-term dependencies in a way that is not feasible with standard convolutional or recurrent operations.

4

4.2.2 REGION ATTENTION

The best attempt of weighted averaging approach that could go under the name of Region Attention, to the best of our knowledge, has been done by [1], who believe that a good pooling or aggregation strategy should adaptively weigh and combine the information across all parts of multimedia content. Their Neural Aggregation Networks (NAN) served as a general framework for learning content-adaptive pooling, emphasizing or suppressing input elements via weighted averaging. The concept of Regional Attention as developed in Section 4.3 is a direct evolution of what has been applied on Face Expression Recognition in [21]. The authors built a so-called Region Attention Network (RAN), capable of extracting features from several spatial regions of the original images, and combining them from a weighted perspective. This method is more robust to occlusion and can better attend to the specific face parts that characterize the human expressions.

4.3 METHOD

In our approach, we use the Inflated 3D ConvNet (I3D) [2], optimally fine-tuned for the classification task at hand, as a feature extractor for multiple video regions and timesteps. These representations are fed to a novel attention module, that summarizes them into a compact feature vector. We experiment with two variants of the module: (spatial) Region Attention (RA) and Temporal Attention (TA), used both individually and jointly.

4.3.1 I3D AND REGION ATTENTION

The Region Attention module produces fixed-length representations that highlight the most informative regions received as input. To achieve this, frames are partitioned with an overlapping regular $N \times N$ grid, with $N = 3$, to extract crops. The attention mechanism is built on top of I3D, which processes the raw videos and outputs respective feature representations. The full model can be trained in two steps. To provide coherent features, I3D is fine-tuned on the multiple video regions that will be considered by the attention module. Each video is handled in a fixed mode: *i.* the video frames are converted to RGB and normalized within the range $[-1.0, 1.0]$; *ii.* $T = 64$ timesteps, of 8 consecutive frames each, are uniformly selected from the full clip; *iii.* through a grid-like scheme, R squared spatial regions are cropped from the fixed-length sample, and resized to I3D input's spatial size 224×224 . The resulting region crops are partially overlapped, since the cropping portion is $5/8$ of a frame. $R = 10$ because the full frame is considered together with the 9 grid regions to preserve global information. $R = 11$ when Actor Focus is applied.

During each I3D training epoch, for each video in the training and validation splits one of the spatial regions is randomly selected. First, this provides data augmentation. Second, I3D extracts features according to the region given as input, instead of always seeing a full frame, thus learning the importance of details in different locations and scale. This behaviour is consistent with the following Region Attention module, that learns to weight the region features, thus making I3D a suitable backbone. Within the Region Attention module a weight in $[0.0, 1.0]$ is assigned to each region feature, through a shared fully-connected layer + Sigmoid activation. The values are used to compute a weighted average of the features, unweighted on the temporal dimension, which is fed into a classification layer. The full process is shown in Figure 4.1.

4.3.2 TEMPORAL ATTENTION

A similar scoring mechanism can be applied to the timesteps. The idea of using attention in the temporal dimension derives, for example, from the fact that initial frames generally have a relatively lower relevance compared to the last frames, which show the result of the activity. Also, in some timesteps the activity does not happen at all. However, extended ablation studies showed that Temporal Attention loses its effectiveness when I3D is fine-tuned, as it appears that the I3D model collects already sufficient information from the sequence of the timesteps. Finally, assuming independence between region importance and timesteps importance,

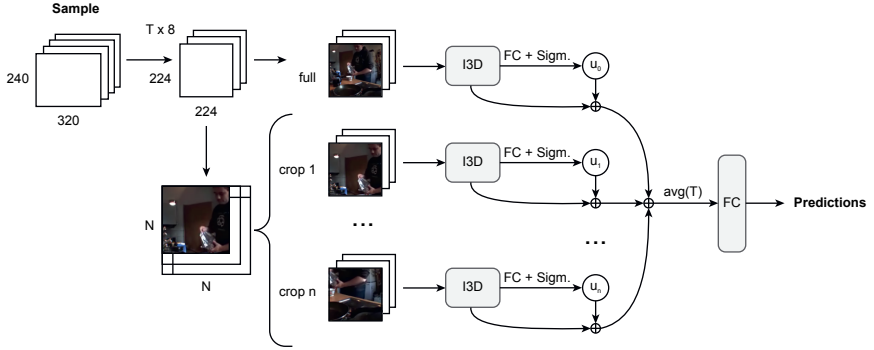


Figure 4.1: The Region Attention module. From every sample in the dataset a 3 x 3 grid is used, and the extracted crops are placed next to the full frames for I3D feature extraction. A fully-connected (FC) layer and a Sigmoid function attribute to each region a score, through which the features are averaged in a weighted manner and feed the final classification layer.

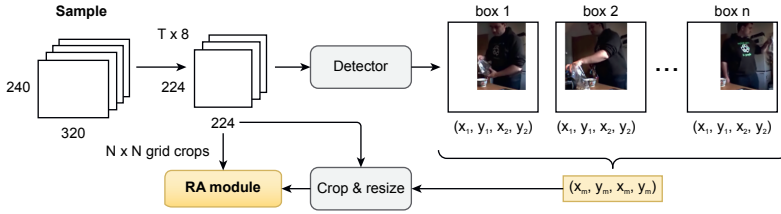


Figure 4.2: The Actor-Focused crop selection through person detection in video frames. Bounding box coordinates for the actor detected in each frame are averaged and used to crop the original video around the person performing the activity. The selected region is added to the others to feed the Region Attention module.

we explored the integration of Region Attention and Temporal Attention by using concatenation, as shown in Figure 4.3.

4.3.3 ACTOR-FOCUS

A further improvement is driven by the consideration that in a high number of cases a single person is performing the activity, generally in a static spatial region of the video. Person detection finds its utility here for the action classification task,

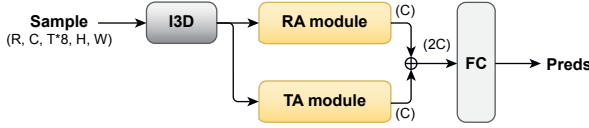


Figure 4.3: Concatenation of regional and temporal features of a video for classification. The two feature vectors computed separately from the two modules are concatenated along the channels and feed the final classification layer.

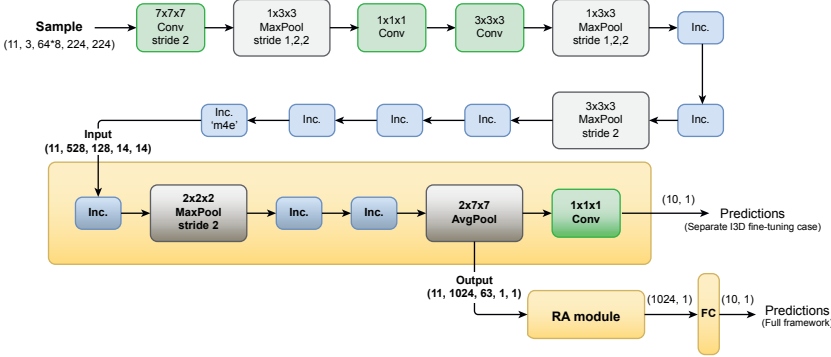


Figure 4.4: I3D + RA architecture. The fine-tuned section (last 3 Inception blocks), together with the RA module and the classification layer, composes the trainable part of the framework, highlighted in yellow. Note that the 1x1x1 Convolution, used as a fully-connected layer in the original I3D architecture, is not used when extracting the features from fine-tuned I3D.

due to the following: *i.* detecting the people in the scene allows the focus to be on the subject performing the activity and on the closest involved entities; *ii.* I3D fine-tuning can be carried out exploiting spatial crops centered on the actor, additionally boosting the ability of the framework to prioritize and highlight the activity globe against clutter and irrelevant background.

For each video, FacebookAI’s Detectron2 [22] is used to get the person bounding box from each frame. As shown in Figure 4.2, the coordinates are averaged, images are cropped accordingly and then resized. Specifically, a square with the same center of the average bounding box and dimensions equal to the biggest between height and width of the bounding box is taken. Since the person box has almost always a higher value for the height than for the width, this means that de-

spite the process of having a fixed averaged bounding box across the video, the actor is likely not to be cut out of the scene when performing small movements. These actor-centered videos are fed to I3D and Region Attention together with the other regions coming from the fixed-grid selection.

4.4 EXPERIMENTS

4.4.1 DATASET

The Breakfast Actions Dataset [7], on which we achieve state-of-the-art results, comprises 10 classes of long-term activities performed by 52 actors. Videos of the first 44 actors are used for training, the remaining for testing. We keep 5 actors from the training split for validation. This gives, respectively, 1322, 411 and 256 videos. The up-to-10-minutes long videos (2 minutes on average) are handled to be of fixed length and size as explained in Section 4.3. To obtain equal width and height the horizontal central crop of each original frame is resized and considered as the selected frame. The resulting frames feed both the grid-like region selection and the person detection mechanism.

4.4.2 ACTOR-FOCUSED I3D + RA

The full Actor-Focused I3D + RA model, unless otherwise specified, considers 11 regions in total. These include the full frame, kept in order to preserve information about the global spatial context from which the regions are extracted, and the actor-centered region. The original I3D implementation remains unchanged except for the very last layer, which is newly initialized considering a 10-fold output due to the number of Breakfast classes. This allows for the utilization of pre-trained I3D checkpoints obtained from Kinetics 400 [2].

Experiments were run on Nvidia GeForce GTX 1080 and Tesla V100 GPUs. Due to the large size of the input and the huge number of parameters of I3D (tens of millions), the devices capacity enabled a maximum batch size of 4 for the backbone fine-tuning. In addition, to make the computation feasible, we restrict the fine-tuning only to the last three Inception blocks and freeze the bottom layers. The features processed by I3D are extracted from the $2 \times 7 \times 7$ AvgPool layer, and feed the conclusive RA step. Again, RA calculates importance scores for each input region and uses them in a weighted average, to aggregate the multi-regional input in a compact representation. The output is a 1024-dimensional vector (2048 in the Region + Temporal Attention setting) and is used for the final classification

step. The full architecture, detailed on input and output shapes, is shown in Figure 4.4.

The developed framework is implemented using PyTorch and trained on single GPU for 100 epochs, using Adam optimizer with learning rate 10^{-3} , ϵ value 10^{-8} , weight decay coefficient 10^{-5} , and CrossEntropy loss function calculated on the 10-fold logits of the last fully-connected layer. Results are calculated on the test set, while our best models are chosen based on the best validation accuracy obtained in 100 epochs.

4.4.3 ABLATION STUDIES

TEMPORAL DIMENSION

First, we show that the amount of timesteps considered has a remarkable impact on classification. Consequently, we confute the assumption that only a few specific moments in time are sufficient for the classification of complex activities. Previous work [4] shows that a uniform selection works generally better than sampling timesteps randomly. Therefore, we keep this setup, and vary instead the quantity of input timesteps, from 4 to 128. Each timestep is composed of 8 consecutive frames.

T	4	16	64	128
Acc. %	68.13	83.94	89.84	86.13

Table 4.1: Full framework results varying the timestep number. Best accuracy on the test set has been reached with $T = 64$.

The results, shown in Table 4.1, indicate that, for an accurate classification, a sufficiently but not exceedingly high number of timesteps from the videos should be considered. This finding is coherent with the complex and variable nature of long-term activities, that are characterized by the presence of several unit-actions. The unit-actions should be represented by the selected video timesteps. Also, sampling a large amount of timesteps helps reduce the noise in input signals, leading to a more robust modelling of the underlying features. However, the results show that an excessively long input might not be optimal. In fact, the highest accuracy obtained with our full model (89.84%) is achieved with $T = 64$, while the accuracy drops when using 128 timesteps. This unexpected outcome can be motivated by considering that many videos in Breakfast are shorter than 128×8 frames = 1024 frames. In this short videos, the 128 selected segments significantly overlap, thus introducing high redundancy and altering the temporal dynamics.

	I3D		I3D + RA	
	val. acc. %	test acc. %	val. acc. %	test acc. %
512 equally spaced frames	83.59	80.05	87.89	86.86
T = 64 (8 frames each)	82.03	82.97	87.50	89.84

Table 4.2: Comparison between frame filtering methods on validation and test sets. Despite a lower accuracy in validation, selecting uniformly 64 timesteps from each video gives better results on the test set. Here, I3D is fine-tuned according to the Multi-Regional with Actor-Focus setting.

4

Following the analysis on the number of video timesteps, we demonstrate that the overall temporal order of the timesteps carries valuable information. First, we shuffle the timesteps during the I3D fine-tuning. As convolution is not a permutation invariant operator, the shuffling has a negative impact on the backbone, and consequently on the Region Attention. With this setup, we obtain an accuracy of 79.81%. We report the results in Table 4.3, under "Sh. timesteps".

Second, we investigate two methods for the feature extraction, that are allowed by the peculiar architecture of I3D. Specifically, thanks to the cascading layers containing max pooling, I3D shrinks the temporal dimension of the input of a factor. As each timestep is composed of 8 consecutive frames, the output feature representation has the same length as the number of timesteps. Because of this, it is possible to extract the features one timestep at a time (*One-at-a-time*) and concatenate the results on the time dimension dimension, or to feed in input all the segments together (*One-shot* fashion), without changing the output size. The difference between the two settings is given by the fact that in the *One-at-a-time* case, the modelling of one specific timestep is not affected by the neighbouring timesteps. On the other hand, in the *One-shot* way the full I3D temporal receptive field is exploited, combining local with global information.

Experiments show that the *One-shot* setting brings a noticeable improvement over *One-at-a-time* features. Intuitively, considering the context in which timesteps are placed, helps achieve a better feature representation. The results from these two setting, respectively, are 89.84% versus 83.7%, as shown in Table 4.3.

The variability in length of Breakfast videos, also within the same class, makes it challenging to represent all the videos fairly in a fixed-length vector. To this extent, short videos are well represented by $T = 64$ timesteps, but this amount of timesteps might not be enough to cover all the unit-actions in longer videos. Other than the uniform and random 8-frame timestep selection evaluated in previous work [4], we experiment with 512 equally spaced frames (*One-shot + 512 f.*) in Table 4.3. Despite achieving slightly better performance in validation (Table 4.2), the *One-shot + 512 f.* setup results in lower accuracy on the test set. This is probably due

to the fact that sampling equidistant frames introduces variable frame frequency in the I3D input. Opposite to this, when sampling timesteps instead of frames, the frequency within each timestep is fixed, as all the videos have the same frame rate. The variable frame frequency alters the motion dynamics modeled by I3D, making the learning process harder.

The last experiment with regards to the temporal dimension is about Temporal Attention, used as an alternative of spatial Region Attention or in conjunction with it. As shown in Table 4.3, applying TA and TRA (combined Temporal-Region Attention, as described in Section 4.3) on top of the convolutional backbone does not result in interesting improvements. Apparently, I3D itself learns sufficiently strong fine-grained and long-term temporal patterns in the fine-tuning phase, thus making Temporal Attention superfluous. On the other side, it is interesting to note that without fine-tuning I3D, the best performances are given by the combination of Temporal and Region Attention. All the above results are summarised in Table 4.3.

I3D setting	T	Acc.	Top	Acc.
Not fine-tuned	64	58.88	TA	65.94
			RA	69.59
			TRA	71.53
One-shot	64	82.97	TA	84.67
			RA	89.84
			TRA	86.62
Sh. timesteps	64	73.97	RA	79.81
One-at-a-time	64	77.62	RA	83.70
One-shot	512 f.	80.05	RA	86.86

Table 4.3: Ablation results considering the temporal axis. Table sections from the top: i. Region Attention (RA), Temporal Attention (TA), Temporal-Region Attention (TRA) on top of not fine-tuned I3D; ii. RA/TA/TRA on top of fine-tuned I3D; iii. same of ii. with different input settings.

SPATIAL DIMENSION

Having discussed the experiments on the temporal axis, we now analyse the spatial dimension. In the following experiments we compare our full model with two model variations: *i.* a simple *Region Mean* model processes 11 video regions and computes a compact representation by taking the arithmetic mean of the features, neglecting the variable importance of the video regions; *ii.* the multi-regional fine-tuning strategy for I3D is replaced with a single region, that corresponds to the

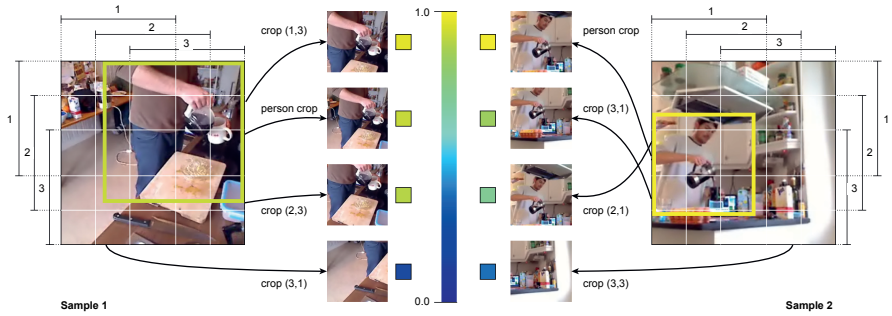


Figure 4.5: Visualization of the different scores that the Region Attention module attributes to the video regions. The four regions that are visualized correspond, respectively, to the top-3 and last crops, for 2 samples of the activity “preparing coffee”. The coloured square in each frame represents the Actor-Focus region. The RA module sets higher scores for the person-centered and grid-central crops.

4

person-centered crop in each training video. To this end, we exploit the Actor-Focus mechanism described in Section 4.3.

The first setting aims to show the improvements brought by RA scoring mechanism. Without the weighted average, the drop in accuracy is around 1.76%, as shown in Table 4.4 (*Region Mean* versus *RA*). Secondly, the comparison with the one region I3D fine-tuning proves the benefit of the multi-regional setup. In fact, training the network with multiple region crops from the same videos acts as a convenient data augmentation strategy. In addition, this learning process produces spatio-temporal features that are more representative of what the following Region Attention module expects as input. When fine-tuning the backbone only with the Actor-Focus crop, the accuracy is 86.62%, with a drop of 3.22% compared to the Multi-Regional setup, as shown in Table 4.4.

Figure 4.5 provides a visualization of the variable importance scores attributed to different video regions through the attention mechanism. According to the prior assumption that the regions of interest for activity recognition revolve around the actor performing the action, RA assigns the highest scores to the person-centered and central grid crops. On the contrary, background regions such as lower and “corner” crops score weights that are close to zero.

Finally, we measure the benefit brought by the Actor-Focus mechanism. The model is trained with and without the Actor-Focus crop. The inclusion of the latter region appears to have a huge impact in the action recognition performance, that increases from 86.62% (*MR I3D* setting in Table 4.4) to the final result of 89.84%.

Backbone	R	Acc.	RA setting	Acc.
I3D not f.t.	1	58.88	RA	72.02
I3D	1	80.05	RA	83.45
MR I3D	10	81.02	RA	86.62
AF I3D	1	81.75	RA	86.62
AF MR I3D	11	82.97	Region mean RA	88.08 89.84
I3D full f.t.	1	80.64	ActionVlad	82.67
			Nonlocal	83.79
			Timeception	86.93
			PIC	89.84

Table 4.4: Ablation results considering the spatial axis. Table sections from the top: i. different I3D fine-tuning settings and Region Attention (RA); ii. best I3D model with Region Mean or RA; iii. former state-of-the-art results on Breakfast. Note: "MR I3D" indicates Multi-Regional fine-tuning on 10 regions (no person-centered region), while "AF I3D" indicates fine-tuning only on person-centered region. R specifies the number of regions. The RA setting is intended to be placed on top of the respective I3D setting.

I3D FINE-TUNING

The extensive experimental comparison between current state-of-the-art methods, is partially limited by the lack of hardware resources. In all the above experiments, I3D is fine-tuned only in the last three convolutional layers and only one region at a time is fed for each video. We leave the end-to-end training of the full Multi-Regional I3D + RA for future work. However, the classification accuracies achieved when fine-tuning the last three layers of I3D or the full model are nearly equal. Respectively, these correspond to 80.05% and 80.64% [5]. As the difference is not significant, we do not expect substantial improvements with a full fine-tuning.

4.5 CONCLUSION

We introduce Multi-Regional I3D fine-tuning with Actor-Focused Region Attention, a neural framework dedicated to the spatio-temporal modelling of long-term activities in videos. We show that the model can learn long-term dependencies across timesteps, resulting in robust representations, and that it is not possible to accurately classify long activities from a few timesteps only. We give insights on the amount of timesteps, their order and the importance of the frame frequency.

Next, a Region Attention module supports spatio-temporal data to adaptively learn the importance of the spatial cues in different video regions, which also allow the backbone to learn rich feature representations. Lastly, an Actor-Focus mechanism drives the attention on the truly discriminative video regions where the actor is performing the activity, neglecting background and irrelevant regions. We demonstrate the effectiveness of the architecture, benchmarking our model on the Breakfast Actions Dataset, with a SOTA-matching accuracy of 89.84%. Because of the modularity of our architecture and of related work [3–5], our framework could complement other approaches. Due to the fact that the strength of our model relies on the way the backbone is fine-tuned and on the use of attention to account for the spatial dimension, further modelling of the time dimension could improve the results. Both PIC [5] and Timeception [3] successfully exploit the time axis and can be juxtaposed on existing backbones, integrated with our RA module. Experiments are left for future work. Finally, future work may include studies on the full I3D fine-tuning and on a I3D + Region Attention end-to-end training.

REFERENCES

- [1] J. Yang, P. Ren, D. Zhang, *et al.* “Neural aggregation network for video face recognition”. In: *CVPR*. 2017.
- [2] J. Carreira and A. Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset”. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.
- [3] N. Hussein, E. Gavves, and A. W. Smeulders. “Timeception for complex action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 254–263.
- [4] N. Hussein, E. Gavves, and A. W. Smeulders. “Videograph: Recognizing minutes-long human activities in videos”. In: *arXiv preprint arXiv:1905.05143* (2019).
- [5] N. Hussein, E. Gavves, and A. W. Smeulders. “Pic: Permutation invariant convolution for recognizing long-range activities”. In: *arXiv preprint arXiv:2003.08275* (2020).
- [6] X. Wang, R. Girshick, A. Gupta, and K. He. “Non-local neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7794–7803.
- [7] H. Kuehne, A. Arslan, and T. Serre. “The language of actions: Recovering the syntax and semantics of goal-directed human activities”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 780–787.
- [8] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. “Hollywood in homes: Crowdsourcing data collection for activity understanding”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 510–526.
- [9] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. “Every moment counts: Dense detailed labeling of actions in complex videos”. In: *International Journal of Computer Vision* 126 (2018), pp. 375–389.
- [10] M. Kalfaoglu, S. Kalkan, and A. Alatan. “Late temporal modeling in 3D CNN architectures with Bert for action recognition”. In: *arXiv*. 2020.
- [11] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei. “Learning spatio-temporal representation with local and global diffusion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12056–12065.
- [12] K. Schindler and L. V. Gool. “Action snippets: How many frames does human action recognition require?” In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

- [13] G. Varol, I. Laptev, and C. Schmid. “Long-term temporal convolutions for action recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017.
- [14] C. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krähenbühl, and R. B. Girshick. “Long-Term Feature Banks for Detailed Video Understanding”. In: *CoRR* abs/1812.05038 (2018). arXiv: 1812.05038. URL: <http://arxiv.org/abs/1812.05038>.
- [15] G. J. Burghouts and K. Schutte. “Spatio-temporal layout of human actions for improved bag-of-words action detection”. In: *Pattern Recognition Letters*. 2013.
- [16] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. “Temporal relational reasoning in videos”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 803–818.
- [17] N. Hussein, M. Jain, and B. E. Bejnordi. “Timegate: Conditional gating of segments in long-range activities”. In: *arXiv preprint arXiv:2004.01808* (2020).
- [18] X. Wang and A. Gupta. “Videos as space-time region graphs”. In: *ECCV*. 2018.
- [19] R. Herzig, E. Levi, H. Xu, *et al.* “Spatio-temporal action graph networks”. In: *ICCV*. 2019.
- [20] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. “Actionvlad: Learning spatio-temporal aggregation for action classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 971–980.
- [21] K. Wang, X. Peng, J. Yang, *et al.* “Region attention networks for pose and occlusion robust facial expression recognition”. In: *IEEE Transactions on Image Processing*. 2020.
- [22] Y. Wu, A. Kirillov, F. Massa, W. Lo, and R. Girshick. *Detectron2*. <https://github.com/facebookresearch/detectron2>. <https://github.com/facebookresearch/detectron2>. 2019.

5

ARE CURRENT LONG-TERM VIDEO UNDERSTANDING DATASETS LONG-TERM?

Many real-world applications, from sport analysis to surveillance, benefit from automatic long-term action recognition. In the current deep learning paradigm for automatic action recognition, it is imperative that models are trained and tested on datasets and tasks that evaluate if such models actually learn and reason over long-term information. In this work, we propose a method to evaluate how suitable a video dataset is to evaluate models for long-term action recognition. To this end, we define a long-term action as excluding all the videos that can be correctly recognized using solely short-term information. We test this definition on existing long-term classification tasks on three popular real-world datasets, namely Breakfast, CrossTask and LVU, to determine if these datasets are truly evaluating long-term recognition. Our study reveals that these datasets can be effectively solved using shortcuts based on short-term information. Following this finding, we encourage long-term action recognition researchers to make use of datasets that need long-term information to be solved.

This chapter has been published as:

O. Strafforello, K. Schutte and J. C. van Gemert. “Are current long-term video understanding datasets long-term?”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2023, pp. 2967-2976

Code available at:

https://github.com/ombretta/longterm_datasets

5.1 INTRODUCTION

Many interesting actions happening in the real world are long-term. That is, they are composed of several short sub-actions, that we refer to as *short-term actions*. For an action to be *long-term*, we deem that recognizing a single-short term action is not enough, and reasoning about the order and the relationship of short-term actions is required. Two examples of long-term actions, shown in Figure 5.1, are *winning a soccer game* and *shoplifting in the supermarket*. To understand which team is winning a soccer game, it is necessary to recognize and count the goals scored since the beginning of the game. For the other example, recognizing if a person is shoplifting, it is necessary to observe a person storing a product in their pocket *and* leaving the supermarket without paying. In both examples, it is not possible to recognize the actions without reasoning on multiple ordered short-term actions.



Figure 5.1: Example of truly long-term actions. *Top*: Who is winning this soccer game?¹, *Bottom*: Is this person shoplifting in the supermarket?². In both cases, it is not possible to answer correctly without considering multiple short-term actions together, their order and relations over time. To understand who is winning the soccer game, it is necessary to recognize and count the goals scored since the beginning of the game. To recognize shoplifting, it is not enough to see a person putting a product in their pocket: also the short-term action *leaving without paying* needs to occur.

¹Source: YouTube; ²Source: YouTube from movie *Un povero ricco*, by Pasquale Festa Campanile (1983).

Achieving automatic long-term action recognition is important because it can

be used to solve real-world problems, from analyzing sports videos, to understanding movies and recognizing threats in surveillance footage. To make it possible, we need purpose-built computer vision models, that are trained and evaluated on datasets that need long-term reasoning to be solved. While working on long-term action recognition, we notice that every video in the Breakfast dataset [1], a go-to choice in long-term video understanding research [2–5], contains short-term actions that map to a single long-term action. This implies that accurately recognizing a short-term action in a Breakfast video should be sufficient to infer the corresponding long-term action. We analyze the short-term actions of another popular instructional video dataset, CrossTask [6], and find the same occurrence in 97.72% of its primary tasks videos. We illustrate our statistics on the short-term action occurrences in Figure 5.2. Since deep learning models are known to use shortcuts to solve classification tasks [7], the models trained and tested on these datasets might learn to exploit short-term information, without encoding any long-term relations.

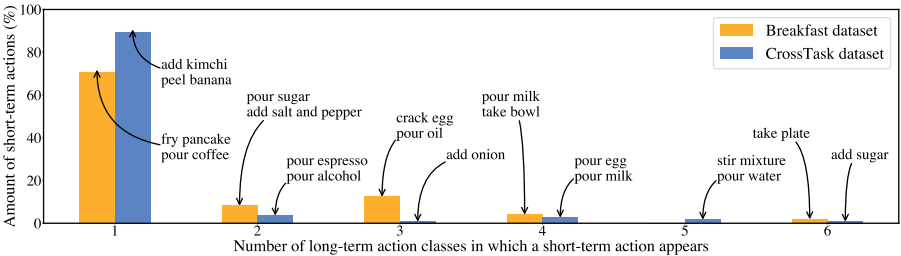


Figure 5.2: We analyze two popular long-term datasets with long-term and short-term action annotations, Breakfast (coarse annotations) [1] and CrossTask [6] (primary tasks). We count in how many long-term actions the short-term action appears. Recurrent short-term actions, like *pour milk* and *pour egg* appear in four different long-term action classes. More specific short-term actions, like *fry pancake* and *add kimchi*, only occur in one long-term action class. We find that a large percentage of short-term actions (70.8% for Breakfast and 89.5% for CrossTask) appears only in one long-term action class. This implies that recognizing a single short-term action might be sufficient to correctly infer the long-term actions in these datasets.

Motivated by this finding, we propose a method to diagnose whether a long-term dataset is suitable to study long-term action recognition, or can be solved using solely short-term information. To this end, we define two requirements for an action to be long-term: (1) The action is *recognizable only from multiple short-term actions* and not from a single short-term action. (2) The action maps to a *single label*. The first requirement makes long-term action recognition impossi-

ble without reasoning over an extended time span. Models that lack this capability, for example based on straightforward pooling operations over time [8], cannot recognize long-term actions. The second requirement leads to discarding multi-label action recognition datasets, like Charades [9], MultiTHUMOS [10] and EPIC-Kitchens [11], as long-term action datasets. In these datasets, the task is to recognize each short-term action contained in the videos. This task could be solved by classifying each short-term action one at a time, while here we are interested in the case where the classification can be made only after reasoning over multiple short-term actions together.

We design a user study to assess whether a video dataset contains long-term action videos that are not recognizable from a single short-term action. Our study is based on two surveys where users have to watch a video and predict the long-term action being performed in the video. In the *Full Videos Survey*, the users can watch the full video, while in the *Video Segments Survey* a separate group of users can watch only a single short clip extracted from the full video. We measure the average action recognition accuracy of the users per video for each survey. The *Full Videos Survey* gives an upper bound to the user long-term action recognition performance. Comparing the accuracy obtained from the *Video Segments Survey* to the upper bound gives an estimate of how many videos in the dataset require long-term information to be correctly recognized. If the action recognition performance of the two groups of users is close, we can conclude that most of the videos in the dataset are not suitable to train and evaluate models for long-term action recognition, because they can be recognized solely by exploiting short-term information.

We apply our proposed method to the aforementioned Breakfast and CrossTask datasets and to the Long-form Video Understanding benchmark (LVU) [12], recently proposed for long-term video recognition tasks in movies. We implement the user studies on Amazon Mechanical Turk [13] and collect responses from more than 150 users. Our results show that looking at a single short video segment is sufficient to recognize 90% and 97.2% of the analyzed videos from Breakfast and CrossTask. Similarly, we find that most of the content understanding tasks in LVU can be classified without long-term information, and that some video segments in this dataset are misclassified by users due to annotation noise. We conclude that the aforementioned datasets might not be suitable to develop new methods for long-term action recognition in videos, because they can be solved by ignoring long-term information. We recommend long-term video understanding researchers to be careful when using these datasets and encourage the community to collect more representative video datasets.

In summary, the contributions of our study can be outlined as follows: (1) We provide a definition of long-term action datasets that should prevent long-term

action recognition models to use traditional short-term action recognition as a shortcut to solve the task. (2) We introduce a method to investigate whether a video dataset meets this definition of long-term action. (3) We find that short-term information is, in most cases, sufficient to solve long-term video understanding tasks in three commonly used datasets. Thus, we recommend against using these datasets in further research on long term action recognition models. The code and responses from our user study are publicly available¹.

5.2 RELATED WORK

5.2.1 ACTION RECOGNITION WITH DEEP LEARNING

The progress of deep learning (DL) has brought significant advancements in automatic action recognition. DL-based models learn to extract discriminative spatial and temporal features directly from the RGB frames of the training videos. Current action recognition models are composed of 3D convolutional networks [14], like I3D [15], C3D [16], Slow-Fast [17]. More recently, attention-based architectures have also shown competitive performance on action recognition tasks. Examples include ViViT [18], TimeSformer [19] and Video Swin Transformer [20]. When pre-trained on sufficiently large datasets, like Kinetics [15] or ActivityNet [21], these models can achieve state-of-the-art action recognition on *short* videos datasets, like UCF101 [22], HMDB51 [23] and Something-Something [24]. However, they are not suitable to learn long-term dynamics in long videos, either due to their limited temporal receptive field or the high computational requirements.

5.2.2 LONG-TERM ACTION RECOGNITION

Long-term action recognition refers to the task of recognizing and understanding human actions composed of several short-term actions, possibly involving multiple objects and movements [5]. Examples include cooking a recipe [1], performing a medical surgery [25] or playing a sport game [10]. Usually, long-term actions require an extended period of time to be executed, e.g. above one minute [3]. Several works that tackled the problem of long-term action recognition use different names and definitions for the same concepts. In fact, long-term actions can also be referred to in the literature as *long-range activities* [26] or *complex activities* [2, 3]. Being composed of multiple steps, the activities in *instructional videos* share the same properties of long-term actions [4, 27, 28] and can be comprised into

¹https://github.com/ombretta/longterm_datasets

this category. Finally, also *long-form* video understanding involves reasoning over human-object interactions in long videos [12, 29] and can be considered as an instance of long-term action recognition.

Traditional DL-based action recognition models [8, 15–17] are deemed insufficient to capture discriminative spatio-temporal features that encode long-term information and the semantic relations between the sub-actions. A variety of models have been proposed to overcome this limitation. Hussein *et al.* [3] proposed to capture long-term information with multi-scale temporal convolution. Yu *et al.* [30] used Recurrent Neural Networks to model long video sequences capturing temporal information at different rhythms. Ballan *et al.* [31] showed that explicitly focusing on the actor performing the long-term action improves the recognition performance. Different approaches showed that long-term action recognition can be tackled using graph-based representations, where the nodes correspond to short-term entities and the edges to their interaction over space and time [5, 32, 33]. Finally, Transformer architectures have been designed to model long-term information in a compute- [34, 35] and data-efficient [2] fashion.

Despite their success, DL-based action recognition models can find shortcuts in the data that let them solve action recognition without learning semantic features, for example classifying the action based on the background scene [7, 36, 37]. In this work, we try to address this problems by analyzing whether commonly used video datasets for long-term action recognition are representative for training DL models, or can be solved using short-term shortcuts.

5

5.2.3 LONG-TERM VIDEO DATASETS

Several datasets have been proposed in the literature to study long-term video understanding tasks. CATER [38] is an ideal example of a dataset that requires long-term information. It involves tracking geometrical shapes that move in a 3D space over time. Sometimes bigger shapes incorporate smaller shapes, rendering their localization impossible without continuous reasoning about past information. As a consequence, models that are not truly long-term fail on this dataset. Unfortunately, the CATER dataset is highly synthetic and cannot be used to train models for real-world applications.

Real-world datasets mostly include cooking [1, 11, 28, 39], home activities [9, 40], sports [10] and instructional videos [6, 28, 41, 42]. A comprehensive overview of long-term video understanding datasets is provided in Table 5.1. Many of these datasets, for example Charades [9], Epic Kitchens [11] and MultiTHUMOS [10], contain long videos annotated with fine-grained, short-term actions. They can be used for multi-label action recognition, where the task is to predict every short-

term action occurring in the video, or for fine-grained action localization. Differently, here we are interested in the single-label classification case, where a global label describes the long-term activity happening in the video. The single label should be recognizable only by reasoning over multiple short-term actions.

Previous work showed that video datasets are sometimes biased towards appearance [43] and better recognizable by short-term over long-term information [44]. Similarly, in this work we explore whether the global labels of datasets proposed for long-term video understanding tasks can be predicted without long-term information. We choose for our study three popular datasets that include single, video-level labels and cover different long-term dataset categories: Breakfast, CrossTask and LVU. Breakfast [1] is a *complex action recognition* dataset used in several works on long-term video understanding [2–5]. CrossTask [6] is a dataset of *instructional videos*, which are composed of several short-term steps that contribute to the completion of a long-term task. Finally, the *Long-form Video Understanding* (LVU) dataset [12] was proposed to learn complex long-term relationships, in contrast to short-term patterns, in video clips extracted from movies.

Dataset	#Videos	Length	#L.T.	#S.T.
COFFEE [41]	150	2	5	51
Epic-Kitchens [11]	432	7.5	-	149, 323
Breakfast [1]	2k	2.3	10	48
Composite [45]	212	1-23	41	218
Charades [9]	10k	0.5	-	157
50-Salads [39]	54	6.4	-	17
COIN [42]	11.8k	2.4	180	778
IKEA FA [46]	101	2-4	-	12
DAHLIA [40]	51	39	7	-
LVU - Content understanding [12]	226	1-3	4	-
	1.3k	1-3	5	-
	723	1-3	6	-
Multi-THUMOS [10]	413	3	-	65
YouCookII [28]	2k	5.3	89	-
CrossTask [6]	4.7k	3-6	83	517

Table 5.1: Overview of current real-world datasets proposed for long-term video understanding tasks. We report the (approximate) number of videos, the average video length in minutes, the number of global *long-term* (L.T.) and *short-term* (S.T.) action recognition classes, if it applies.

5.3 ASSESSING LONG-TERM ACTION RECOGNITION DATASETS

5.3.1 USER STUDY

According to our definition, an action is long-term if it cannot be classified from a single short video segment. We design a user study to test whether current long-term video understanding datasets respect this property. Our user study consists of two surveys. In the *Full Videos Survey*, the users are presented with the full-length videos from the datasets. In the *Video Segments Survey*, the users are presented with a short video segment extracted from a full-length video. In both surveys, the users are instructed to watch the video clip and express what action is being performed in the full video, in their opinion. The users are provided with a list of possible actions, which correspond to the classes from the analyzed long-term action datasets, and have to select exactly one action class from the list. We include the additional option "*I am not sure*", to let the users express uncertainty when they are in doubt about which action to select.

From the collected user votes in the *Full Videos Survey* and the *Video Segments Survey*, we calculate and compare the action recognition accuracy. If the users from the two groups perform similarly, we can conclude that the videos do not contain long-term actions, as they can be recognized from single short-term actions comparably well than looking at the full videos. We also calculate the user agreement per survey, measured with Krippendorff's α [47], which gives an indication of how subjective the prediction task is. We expect that the more a video is difficult to classify, the more subjective the choice will be, thus resulting in low agreement.

5.3.2 MEASURING RECOGNITION ACCURACY

From the *Full Videos Survey*, we collect user votes per class for each full-length video. In each full video, we express the votes in percentages ($\%user_votes_v(c)$), which we obtain by dividing the votes per class by the amount of votes collected for the full video. As formalized in Equation 5.1, given \mathcal{C} classes from the evaluated dataset, excluding the *I am not sure* option, we assign to the full video prediction ($pred(v)$) the class voted by the majority of the users. The long-term action recognition accuracy is given by the number of full videos assigned with the correct class over the number of full videos considered in the study for the dataset.

$$pred(v) = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \quad \%user_votes_v(c) \quad (5.1)$$

In the *Video Segments Survey*, we collect user votes for every segment s_v in a full video. Again, for each segment we calculate the percentage of votes per class $\%user_votes(c)$. Then, we extract the full video prediction from the votes of a single segment. To do this, we select the segment s_v^* with highest percentage of votes for a single class, excluding the *I am not sure* option. This approach is formalized in Equation 5.2. In the example in Figure 5.3, the full video is assigned the class *Making scrambled eggs*, which is voted by 86% of users in *Segment 5*, which is the maximum ratio of votes for one class across the video segments. According to our definition, if the full-length video is long-term, there should be no video segments that lead to the right predicted class. The accuracy is given by the number of full videos assigned with the correct label over the number of full videos considered in the study.

$$pred(v) = pred(s_v^*), \quad (5.2)$$

$$\begin{aligned} \text{where } s_v^* &= \underset{s_v \in v}{\operatorname{argmax}} \quad \{\max_{c \in \mathcal{C}} \%user_votes_{s_v}(c)\}, \\ pred(s_v^*) &= \underset{c \in \mathcal{C}}{\operatorname{argmax}} \quad \%user_votes_{s_v^*}(c). \end{aligned}$$

5

5.4 RESULTS

We include in our study a representative dataset from complex action recognition, Breakfast [1], one instructional video dataset, CrossTask [6], and the Long-Form Video Understanding (LVU) dataset [12]. We implement the user study on Amazon Mechanical Turk [13] and collect responses from 167 users. We collect, on average, 12.09 ± 1.62 votes for each video and video segment, which is proved to be a proper amount [48]. Table 5.2 provides an overview of the results from the *Full Videos Survey* and the *Video Segments Survey*, discussed in the following sections.

5.4.1 BREAKFAST

Breakfast [1] is a collection of third-person videos of actors cooking a breakfast recipe, like scrambled eggs, coffee, cereals and milk. Each video has a global label,

What action is being performed in this video? (GT: “Making scrambled eggs”)

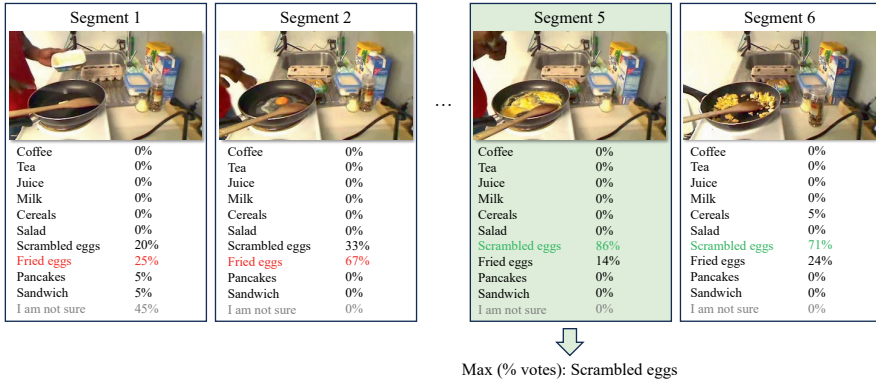


Figure 5.3: In the *Video Segments Survey*, users have to understand what is happening in a long video by looking only at one short segment. We ask the users to vote for a video class and obtain predictions per segment. We assign to the full video the segment prediction with the highest percentage of votes for one class. In the example, taken from the Breakfast dataset [1], *Segment 5* determines the video prediction *Scrambled eggs*.

Dataset	Classification accuracy (%)	
	Full Videos	Video Segments
Breakfast	93.33	90.0
CrossTask	100.0	97.2
LVU – Relationship	88.89	88.89
LVU – Scene	100.0	100.0
LVU – Speaking	80.0	60.0

Table 5.2: Average video recognition accuracy obtained from the *Full Videos Survey* and *Video Segments Survey* on the Breakfast [1], CrossTask [6] and LVU [12] datasets. The results suggest that long-term information is helpful but not necessary in the majority of the evaluated datasets.

which corresponds to the recipe being made, for a total of 10 classes. The classification task consists in correctly recognizing the recipe.

For our study, we select a representative subset of 30 videos, corresponding to 3 randomly selected videos per class. The full videos have average duration of 2.44 ± 2.18 minutes. For the *Video Segments Survey*, we segment the video according to the short-term action timesteps (*coarse segmentation*) provided in the dataset.

Dataset	User agreement		
	Full Videos	Video Segments	Selected Segments
Breakfast	0.717	0.386	0.593
CrossTask	0.671	0.462	0.767
LVU – relationship	0.499	0.340	0.523
LVU – scene	0.755	0.481	0.686
LVU – speaking	0.159	0.191	0.265

Table 5.3: Overview of the user agreement in our user studies, measured terms of Krippendorff’s α [47]. We find that the users tend to agree in the *Full Videos Surveys* and when selecting the segments with highest amount of votes for a class. Recognizing the actions in the *Video Segments Survey* is generally harder then when looking at the full video, resulting in more variability in the users predictions and, consequently, in lower agreement.

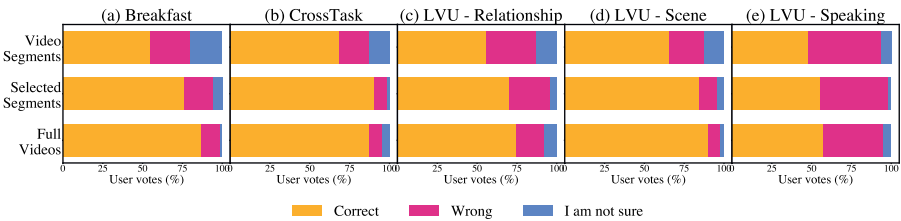


Figure 5.4: Overview of the user votes (correct, wrong and *I am not sure*) collected in our study. We compare the results from the *Full Videos*, all the *Video Segments*, and the *Selected Segments* with highest percentage of votes for one class. The amount of correct votes in the *Selected Segments* is significantly higher than for all the *Video Segments*, and comparable, or even higher, to the amount of correct votes obtained watching the full videos. N.b., the user votes reported in this figure do not have to match the accuracies in Table 5.2. While the accuracy shows the percentage of videos correctly classified, the user votes are aggregated without considering the votes distributions within the specific videos.

We remove segments that are shorter than 5 seconds, as we deem those segments highly uninformative, and we obtain 154 segments in total, of average duration 29 ± 39 seconds, where $\sim 56\%$ of the segments last less than 15 seconds. The large standard deviation is due to some repetitive short-term actions that can last above a minute, e.g. *stir dough* or *fry egg*.

The results in Table 5.2 show that the recognition accuracy from the *Full Videos Survey* (93.33%) and the *Video Segments Survey* (90.0%) are close. This suggests

that, although having access to the full long-term information in the video helps, looking at single short segments is sufficient to infer the right recipe class for the majority of the videos. From this result we conclude that the Breakfast dataset is not a proper long-term action dataset, according to our definition.

We analyze the amount of correct user votes, wrong votes and *I am not sure* votes obtained in the user study and illustrated in Figure 5.4 (a). We obtained 86.78% of correct votes in the *Full Videos Survey* and 54.47% in the *Videos Segments Survey*. However, if we consider only the segments with the highest percentage of votes for one class, the amount of correct votes reaches 76.36%. A similar trend occurs in the user agreement in Table 5.3. By further inspecting the results from the *Video Segments Survey*, we notice that users are generally more uncertain classifying the video segments early in the video, with a higher portion of *I am not sure* votes compare to the later segments. In particular, 63.57% of *I am not sure* votes are obtained in from the first two video segments in chronological order. We argue that breakfast dishes are usually better recognizable towards the end of the video, when the recipe is complete.

5

5.4.2 CROSSTASK

CrossTask [6] is an instructional video dataset of $\sim 4.7k$ videos, covering themes like auto repair, cooking and DIY. The instructional videos show how to perform a *tasks* (e.g., *Make a Latte*) through a list of *steps* (e.g., *add coffee*, *press coffee*, *pour water*, *pour espresso*, *steam milk*, *pour milk*). It contains 18 primary tasks with steps annotations and 65 related tasks with unlabeled steps. The dataset is meant to be used to learn steps in a weakly supervised learning setup. Here, we evaluate whether predicting the *task* illustrated in an instructional video also fits our definition of long-term action recognition. We collect results from 36 video clips (2 random videos per primary task) of average duration 4.50 ± 2.14 minutes. Similarly to Breakfast, we extract 260 segments from the videos according to the timesteps provided with the dataset. In CrossTask, the segments are significantly shorter than Breakfast, with average duration of 10 ± 11 seconds and $\sim 81\%$ of the segments being shorter than 15 seconds.

In Table 5.2, we compare the task recognition accuracy from the *Full Videos Survey*, 100%, and the *Video Segments Survey*, 97.2%. In both cases, users can recognize the task with high accuracy. Only one video (YouTube id *kReUYklvjnc*) is misclassified in the *Video Segments Survey*, despite 5/8 of its video segments being correctly classified. Considering the user agreement (Table 5.3) and correct votes by the users (Figure 5.4, b), we find that both quantities are marginally higher in the Selected Segments over the Full Videos. This result shows that users tend to

make the same mistakes (as for video *kReUYklvjnc*) while confirming that most of the tasks are generally recognizable both from short video segments and full videos. It is worth noting that the results reported in Table 5.2 and Figure 5.4 are not necessarily the same. The accuracy corresponds to the percentage of videos correctly classified, while the user votes are aggregated without considering the votes distributions within the specific videos. Because of the high task recognition accuracy obtained from the *Video Segments Survey*, we conclude that the videos in CrossTask do not contain long-term actions. We recommend to use this dataset for the other video understanding tasks that is supports, like captioning and action localization.

5.4.3 LVU

The Long-Form Video Dataset (LVU) [12] has been recently proposed to study complex relationships in video clips extracted from movies. It provides three tasks, related to content understanding, user engagement prediction and movie metadata prediction and contains over 11k videos. Similarly to previous work [49], we select the task of *Content Understanding*, which involves classifying the *relationship* among the characters, where the *scene* is taking place and the characters *speaking* style, from video clips of ~2.5 minutes. The respective annotations consist in a global label per video. We assess whether predicting *Relationship*, *Scene* and *Speaking* is a form of long-term action recognition, according to our definition. We select videos from the test set and manually extract segments for each of the three classification tasks. We obtain 9 videos (3 per class) for *Relationship*, 12 videos (2 per class) for *Scene* and 10 videos (2 per class) for *Speaking*, and a total of 140 segments of ~30 seconds.

Table 5.2 shows the classification accuracies obtained from the *Full Videos Survey* and *Video Segments Survey*. Comparing the results, we find no difference for *Relationship* and *Scene*. In particular, *Scene* classification is performed with 100% accuracy, indicating that this prediction task is easy for humans. We identify a problem associated with LVU - *Relationship*. The labels husband-wife, friends, boyfriend-girlfriend are associated with specific characters in the movie, but other characters might appear within the same video clip. For example, in Figure 5.5 (a), the ground-truth label for the movie in the first row is *Husband-Wife*. However, a third male character appears in the scene in addition to the *husband and wife*. Therefore, the labels only correctly apply to a specific subset of the characters in the scene, or to a precise time window when only the target characters appear. As a result, the full videos are classified with a high percentage of wrong votes, while some of the video segments that do not include the characters corresponding to

the label are completely misclassified. This justifies the large portion of wrong votes in Figure 5.4 (c) and relatively low agreement in Table 5.3.



Figure 5.5: Examples of correct (green) and wrong (red) classification results collected from the *Video Segments* (V.S.) and *Full Videos* (F.V.) surveys on the Long-form Video Understanding (LVU) - Relationship (a), Scene (b) and Speaking (c) dataset [12]. Users correctly classify a large portion of video segments. Other segments result misclassified due to annotation noise.

We find a similar annotation problem in LVU - *Speaking*. Also in this case, the global label only applies to a subset of the characters in the scene. In the example in Figure 5.5 (c), the label *Threatens* only applies to the man with the gun. This ex-

plains the difference in performance when comparing the accuracies from the *Full Videos Survey* and *Video Segments Survey* in Table 5.2, the large amount of wrong votes in Figure 5.4 (e) and low agreement in Table 5.3. Because of the problem with the annotations and the equal recognition performance of 88.89% obtained from the *Full Videos Survey* and *Video Segments Survey* (reported in Table 5.2), we conclude that LVU - *Relationship* is not a long-term video understanding task. Similar conclusions apply for LVU - *Scene*, with perfect classification scores resulting from both surveys. Finally, the labels in LVU - *Speaking* are not truly long-term, as they apply to a subset of characters speaking only during some relatively short time-windows.

5.5 CONCLUSION

We propose a method to assess whether an action is *long-term*. We apply our method to three current long-term video understanding datasets, Breakfast, CrossTask and LVU. Our results show that long-term information might help but is *not necessary* in the majority of videos from the analyzed datasets. In fact, the long-term actions in these videos can be correctly classified by humans by looking solely at a single short video segment. This result suggests that deep learning models trained and tested on these datasets might pick short-term shortcuts and still show correct recognition performance, without actually learning any long-term information. Following our findings, we urge researchers who are investigating automatic long-term action recognition to use datasets that need long-term information to be solved.

Acknowledgements. This work is part of the research program Efficient Deep Learning (EDL), which is (partly) financed by the Dutch Research Council (NWO).

REFERENCES

- [1] H. Kuehne, A. Arslan, and T. Serre. “The language of actions: Recovering the syntax and semantics of goal-directed human activities”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 780–787.
- [2] H. Guo, H. Wang, and Q. Ji. “Uncertainty-guided probabilistic transformer for complex action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20052–20061.
- [3] N. Hussein, E. Gavves, and A. W. Smeulders. “Timeception for complex action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 254–263.
- [4] M. Li, L. Chen, Y. Duan, Z. Hu, J. Feng, J. Zhou, and J. Lu. “Bridge-prompt: Towards ordinal action understanding in instructional videos”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19880–19889.
- [5] J. Zhou, K.-Y. Lin, H. Li, and W.-S. Zheng. “Graph-based high-order relation modeling for long-term action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8984–8993.
- [6] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic. “Cross-task weakly supervised learning from instructional videos”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3537–3545.
- [7] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.
- [8] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. “Temporal segment networks: Towards good practices for deep action recognition”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 20–36.
- [9] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. “Hollywood in homes: Crowdsourcing data collection for activity understanding”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 510–526.
- [10] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. “Every moment counts: Dense detailed labeling of actions in complex videos”. In: *International Journal of Computer Vision* 126 (2018), pp. 375–389.

- [11] D. Damen, H. Dougherty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.* "Scaling egocentric vision: The epic-kitchens dataset". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 720–736.
- [12] C.-Y. Wu and P. Krahenbuhl. "Towards long-form video understanding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1884–1894.
- [13] *Amazon Mechanical Turk*. <https://www.mturk.com/>. Accessed: 2023-07-05.
- [14] S. Ji, W. Xu, M. Yang, and K. Yu. "3D convolutional neural networks for human action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 221–231.
- [15] J. Carreira and A. Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.
- [16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4489–4497.
- [17] C. Feichtenhofer, H. Fan, J. Malik, and K. He. "Slowfast networks for video recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6202–6211.
- [18] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. "Vivit: A video vision transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6836–6846.
- [19] G. Bertasius, H. Wang, and L. Torresani. "Is space-time attention all you need for video understanding?" In: *ICML*. Vol. 2. 3. 2021, p. 4.
- [20] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. "Video swin transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3202–3211.
- [21] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. "Activitynet: A large-scale video benchmark for human activity understanding". In: *CVPR*. 2015.
- [22] K. Soomro, A. R. Zamir, and M. Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild". In: *arXiv preprint arXiv:1212.0402* (2012).
- [23] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. "HMDB: a large video database for human motion recognition". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2556–2563.
- [24] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.* "The" something something" video database for learning and evaluating visual common sense". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5842–5850.

- [25] A. Sharghi, H. Haugerud, D. Oh, and O. Mohareri. “Automatic operating room surgical activity recognition for robot-assisted surgery”. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* 23. Springer. 2020, pp. 385–395.
- [26] N. Hussein, M. Jain, and B. E. Bejnordi. “Timegate: Conditional gating of segments in long-range activities”. In: *arXiv preprint arXiv:2004.01808* (2020).
- [27] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. “End-to-end learning of visual representations from uncurated instructional videos”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9879–9889.
- [28] L. Zhou, C. Xu, and J. Corso. “Towards automatic learning of procedures from web instructional videos”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.
- [29] X. Yang, F.-J. Chu, M. Feiszli, R. Goyal, L. Torresani, and D. Tran. “Relational Space-Time Query in Long-Form Videos”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6398–6408.
- [30] T. Yu, Y. Li, and B. Li. “Rhyrnn: Rhythmic rnn for recognizing events in long and complex videos”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 127–144.
- [31] L. Ballan, O. Strafforello, and K. Schutte. “Long-term Behaviour Recognition in Videos with Actor-focused Region Attention.” In: *VISIGRAPP (5: VISAPP)*. 2021, pp. 362–369.
- [32] N. Hussein, E. Gavves, and A. W. Smeulders. “Videograph: Recognizing minutes-long human activities in videos”. In: *arXiv preprint arXiv:1905.05143* (2019).
- [33] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles. “Action genome: Actions as compositions of spatio-temporal scene graphs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10236–10247.
- [34] M. M. Islam and G. Bertasius. “Long movie clip classification with state-space video models”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2022, pp. 87–104.
- [35] C.-Y. Wu, Y. Li, K. Mangalam, H. Fan, B. Xiong, J. Malik, and C. Feichtenhofer. “Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13587–13597.
- [36] J. Choi, C. Gao, J. C. Messou, and J.-B. Huang. “Why can’t i dance in the mall? learning to mitigate scene bias in action recognition”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [37] K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry. “Noise or Signal: The Role of Image Backgrounds in Object Recognition”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=gl3D-xY7wLq>.

- [38] R. Girdhar and D. Ramanan. "CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning". In: *arXiv preprint arXiv:1910.04744* (2019).
- [39] S. Stein and S. J. McKenna. "Combining embedded accelerometers with computer vision for recognizing food preparation activities". In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 2013, pp. 729–738.
- [40] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard. "The DAily Home LIfe Activity Dataset: A High Semantic Activity Dataset for Online Recognition". In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. 2017, pp. 497–504. DOI: 10.1109/FG.2017.67.
- [41] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. "Learning from narrated instruction videos". In: *IEEE transactions on pattern analysis and machine intelligence* 40.9 (2017), pp. 2194–2208.
- [42] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou. "Coin: A large-scale dataset for comprehensive instructional video analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1207–1216.
- [43] P. Byvshv, P. Mettes, and Y. Xiao. "Are 3D convolutional networks inherently biased towards appearance?" In: *Computer Vision and Image Understanding* 220 (2022), p. 103437.
- [44] O. Strafforello, X. Liu, K. Schutte, and J. van Gemert. "Video BagNet: short temporal receptive fields increase robustness in long-term action recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2023.
- [45] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. "Script data for attribute-based recognition of composite activities". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2012, pp. 144–157.
- [46] S. Toyer, A. Cherian, T. Han, and S. Gould. "Human pose forecasting via deep markov models". In: *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE. 2017, pp. 1–8.
- [47] K. Krippendorff. "Computing Krippendorff's alpha-reliability". In: *Departmental Papers (ASC), University of Pennsylvania* (2011).
- [48] A. Carvalho, S. Dimitrov, and K. Larson. "How many crowdsourced workers should a requester hire?" In: *Annals of Mathematics and Artificial Intelligence* 78 (2016), pp. 45–72.
- [49] Y. Sun, H. Xue, R. Song, B. Liu, H. Yang, and J. Fu. "Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning". In: *arXiv preprint arXiv:2210.06031*. 2022.

6

VIDEO BAGNET: SHORT TEMPORAL RECEPTIVE FIELDS INCREASE ROBUSTNESS IN LONG-TERM ACTION RECOGNITION

Previous work on long-term video action recognition relies on deep 3D-convolutional models that have a large temporal receptive field (RF). We argue that these models are not always the best choice for temporal modeling in videos. A large temporal receptive field allows the model to encode the exact sub-action order of a video, which causes a performance decrease when testing videos have a different sub-action order. In this work, we investigate whether we can improve the model robustness to the sub-action order by shrinking the temporal receptive field of action recognition models. For this, we design Video BagNet, a variant of the 3D ResNet-50 model with the temporal receptive field size limited to 1, 9, 17 or 33 frames. We analyze Video BagNet on synthetic and real-world video datasets and experimentally compare models with varying temporal receptive fields. We find that short receptive fields are robust to sub-action order changes, while larger temporal receptive fields are sensitive to the sub-action order.

This chapter has been published as:

O. Strafforello, X. Liu, K. Schutte, and J. C. van Gemert. "Video BagNet: short temporal receptive fields increase robustness in long-term action recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2023, pp. 159-166

Code available at:

<https://github.com/ombretta/videobagnet>

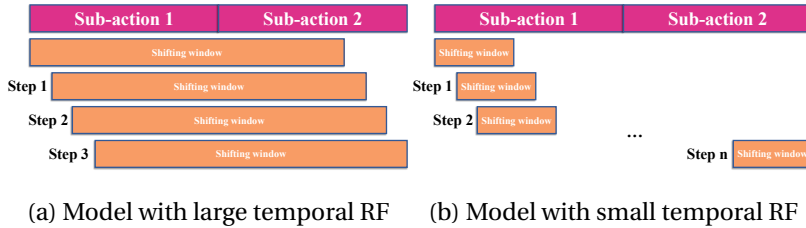


Figure 6.1: Large (a) versus small (b) temporal RF compared to the sub-action duration. The temporal RF size in the last convolutional layer is represented by the size of the convolutional shifting windows. Models with large temporal RF see sub-actions in ordered co-occurrences, while models with small temporal RF are more likely to see single sub-action occurrences. Because of this, models with small temporal RFs encode sub-action occurrences but not strict sub-action orders.

6.1 INTRODUCTION

Long-term action videos naturally have different sub-action combinations and orders. For instance, the action of ‘making coffee’ may contain either order of ‘add sugar, add milk’, or ‘add milk, add sugar’, or people can drink their coffee black. With such diversity in sub-action orders it is nearly impossible to sample representative data containing all possible permutations for training a long-term action recognition classifier. Thus, the training set in current long-term classification datasets like MultiTHUMOS [1] and Charades [2] may contain different sub-action orders than the test set. The specific sub-action order and duration is exploited by current video action recognition models due to their large temporal receptive field size. Consequently, if the models encode the specific sub-action order at training time, it might cause misclassification of a video action when the sub-action order differs at test time.

In this paper, we focus on encoding sub-action *order*. We refer to the *temporal receptive field* (RF) as the number of input frames within a shifting kernel that a network can make use of in its last convolutional layer. Usually, the last convolutional layer is followed by global temporal pooling, which collapses the temporal dimension into one unit, and a final fully connected layer. These operations do not affect the temporal RF size and the sensitivity to order, as they cannot model temporal dependencies. For this reason, we do not consider the final pooling and classification layers in our calculation of the temporal RF size. Networks with temporal RF size larger than the sub-action duration (as shown in Figure 6.1 (a)) might overfit on the exact sub-action order seen at training time. In cases where the avail-

able training samples are not sufficiently representative of all possible sub-action orders, misclassifications occur at test time.

We introduce Video BagNet, a model with a small temporal RF size that is less sensitive to the exact sub-action order. Our model is inspired by BagNet [3], which reduces the spatial receptive field size for easier network interpretation. We use Video BagNet to investigate the role of the temporal RF in encoding the sub-action order. Our proposed Video BagNet is modified from 3D ResNet-50 [4]. We reduce the temporal RF size by shrinking the kernels in the temporal dimension and using less down-sampling. As shown in Figure 6.1 (b), our Video BagNet with small temporal RF sizes is less sensitive to the exact sub-action order by seeing occurrences of single sub-actions rather than the combinations of ordered sub-actions. This results in better sub-action detection performance than 3D ResNet-50 on our synthetic *Directional Moving MNIST* dataset and MultiTHUMOS. We also provide a measurement of model sensitivity to the sub-action order. Our code will be made publicly available¹.

6.2 RELATED WORK

6.2.1 TEMPORAL EXTENT OF RECENT MODELS FOR ACTION RECOGNITION

Recent action recognition architectures can model long temporal extents [5–11]. This is achieved through two main approaches. The first one is by extending the temporal receptive field of convolutional models, either by stacking strided convolutional layers, thus making the model deeper [12, 13], or by harnessing auxiliary temporal modules [5, 9, 14]. The second approach is by means of transformer architectures, whose design entails a temporal receptive field which spans over the whole input duration [15–17]. Large temporal extents make it possible to learn dependencies in videos over time. This allows for modeling the order of the sub-actions that are seen at training time, which is considered useful to capture the inner structure of complex, long-term activities [18].

However, models with large temporal RF have a drawback: they are prone to overfitting on the order when the available training data is limited [19]. This is the case for most of the current long-term action recognition datasets, which only consist of a few hundred or thousand videos [1, 2, 20]. In addition, recent work showed that some of the current long-term action recognition datasets can be solved without using long-term information [21]. In this work, we investigate whether mod-

¹<https://github.com/ombretta/videobagnet>

eling large temporal extents is always beneficial to solve long-term action recognition. In particular, we investigate whether models with large temporal RF overfit on the order of the sub-actions seen at training time, causing misclassifications at test time.

6.2.2 ORDER INVARIANT NETWORKS

In [14], it is empirically shown that the classification performance of order-aware methods drops significantly when new sub-action orders are presented at test time. On the other hand, order invariant methods, like ActionVLAD [22], are robust to sub-actions permutations. Hussein *et al.* [18] propose a permutation invariant convolutional module, PIC, to model temporal dynamics in long-range activities. The PIC module performs self-attention across pre-extracted visual features and can be stacked on top of convolutional backbones. PIC is robust to sub-action permutation compared to ordered-aware convolutional baselines [5], while maintaining a large temporal RF.

Our approach deviates from ActionVLAD and PIC. While ActionVLAD is completely order unaware, we maintain order information within short receptive fields. This allows modeling fine-grained motions, which is proven beneficial for action recognition [23, 24]. Differently than PIC, we investigate sensitivity to sub-action order by looking at the temporal RF size of spatio-temporal convolutional networks, commonly used as backbones in long-term action recognition models [5, 9, 14]. Our method only requires simple modification to the spatio-temporal convolutional networks.

6.2.3 REDUCING THE RECEPTIVE FIELD SIZE: BAGNET

Our idea of reducing the temporal receptive field size is inspired by Brendel *et al.* [3], who investigated how bag-of-local-features can be used for image classification. Bag-of-local-features can be obtained by restricting the spatial receptive field of the image classifier to a small number of pixels. In Brendel *et al.*'s model, the *BagNet*, this is achieved by replacing a set of 3×3 convolutions with 1×1 convolutions and removing the first downsampling layer. The property of this architecture is that the image feature representation is given by a collection of local features, corresponding to small image patches, that do not take into account the global spatial structure. Surprisingly, ignoring global structures does not hurt substantially the classification accuracy of BagNet. Using bag-of-local-features has been taken on for other visual classification tasks. Some examples are exploring local features for face anti-spoofing [25], and predicting the histogram of visual words

of a discretized image as part of a self-supervision task [26]. To the best of our knowledge, our method is the first work that relies on bag-of-temporal-features models to learn video representations.

	3D ResNet-50 (RN)	Video BagNet-1/9/17/33 (BN)	
# parameters for 3 classes	46.2 M	45.9/46.7/45.6/46.5 M	Output sizes $T \times S^2$
conv1	$7 \times 7^2, 64$, stride (1, 2, 2)	$1/3/3/3 \times 7^2, 64 \times k$, stride (1, 2, 2)	RN: 64×32^2 BN: 64×32^2
downsampling	Max pool (3, 3, 3), stride 2	Max pool (1, 3, 3), stride (1, 2, 2)	RN: 32×16^2 BN: 62×16^2
conv2_x	$\begin{bmatrix} 1 \times 1^2, 64 \\ 3 \times 3^2, 64 \\ 1 \times 1^2, 64 \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 256 \\ 3 \times 3^2, 64 \\ 1 \times 1^2, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1^2, 64 \times k \\ 1/3/3/3 \times 3^2, 64 \times k \\ 1 \times 1^2, 64 \times k \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 256 \times k \\ 1/1/1/1 \times 3^2, 64 \times k \\ 1 \times 1^2, 64 \times k \end{bmatrix} \times 2$	RN: 32×16^2 BN: 60×16^2
conv3_x	$\begin{bmatrix} 1 \times 1^2, 256 \\ 3 \times 3^2, 128 \\ 1 \times 1^2, 128 \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 512 \\ 3 \times 3^2, 128 \\ 1 \times 1^2, 128 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1^2, 256 \times k \\ 1/3/3/3 \times 3^2, 128 \times k \\ 1 \times 1^2, 128 \times k \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 512 \times k \\ 1/1/1/1 \times 3^2, 128 \times k \\ 1 \times 1^2, 128 \times k \end{bmatrix} \times 3,$	RN: 16×8^2 BN: 29×8^2
conv4_x	$\begin{bmatrix} 1 \times 1^2, 512 \\ 3 \times 3^2, 256 \\ 1 \times 1^2, 256 \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 1024 \\ 3 \times 3^2, 256 \\ 1 \times 1^2, 256 \end{bmatrix} \times 5$	$\begin{bmatrix} 1 \times 1^2, 512 \times k \\ 1/1/3/3 \times 3^2, 256 \times k \\ 1 \times 1^2, 256 \times k \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 1024 \times k \\ 1/1/1/1 \times 3^2, 256 \times k \\ 1 \times 1^2, 256 \times k \end{bmatrix} \times 5$	RN: 8×4^2 BN: 14×4^2
conv5_x	$\begin{bmatrix} 1 \times 1^2, 1024 \\ 3 \times 3^2, 512 \\ 1 \times 1^2, 512 \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 2048 \\ 3 \times 3^2, 512 \\ 1 \times 1^2, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1^2, 1024 \times k \\ 1/1/1/3 \times 3^2, 512 \times k \\ 1 \times 1^2, 512 \times k \end{bmatrix},$ $\begin{bmatrix} 1 \times 1^2, 2048 \times k \\ 1/1/1/1 \times 3^2, 512 \times k \\ 1 \times 1^2, 512 \times k \end{bmatrix} \times 2$	RN: 4×2^2 BN: 6×2^2
Average pool, n_classes-d fc, softmax			

Table 6.1: Network architectures: 3D ResNet-50 (RN) vs Video BagNet-1, 9, 17 and 33 (BN).

In the first row, we report the number of parameters. The next rows correspond to the network layers, which contain convolutions and downsampling. For the convolutional layers, we report the kernel size $T \times S^2$, in the temporal (T) and spatial (S^2) dimensions, and the number of channels. The rightmost column of the table reports the output sizes at each layer, given an input clip of size 64×64^2 . The convolutional blocks follow the structure of ResNet Bottleneck blocks [27]. We widen the channels of Video BagNet with factor k , equal to 1.40, 1.40, 1.35 and 1.25, to keep the number of parameters comparable among the different models. In both architectures, each layer is followed by Batch Norm [28] and a ReLU [29].

6.3 METHOD

We study how the size of the temporal RF effects model sensitivity to sub-action order. To this end, we compare long-term action recognition performance of 3D convolutional networks with variable temporal RF size.

6.3.1 VIDEO BAGNET

Inspired by the 2D BagNet for image classification [3], we design Video BagNet, a 3D convolutional network that reasons over short temporal extents. The key idea behind Video BagNet is to harness bag-of-feature representations for video classification. Specifically, the word vocabulary is composed of short video segments. Although this representation does not allow to model long-term temporal dependencies, it prevents learning strict temporal orders that can lead to the misclassification of a video if unseen permutations between sub-actions occur at test time.

Our Video BagNet is based on the 3D ResNet-50 described in Hara *et al.* [4]. We apply a set of modifications to 3D ResNet-50 to restrict the size of its temporal receptive field, while leaving the computation in the spatial dimensions unchanged. In particular, we propose four variants of Video BagNet, with temporal RF sizes of 1, 9, 17, and 33 input frames. We choose these temporal extents following the design choice of Brendel *et al.* [3] in the image domain. Video BagNet is sensitive to order within its small temporal RF, allowing for fine-grained motion modeling.

The set of modifications that we apply to 3D ResNet-50 can be summarized as follows.

First, we restrict the size of some of the convolutional kernels in the temporal dimensions. This is done to adaptively control the expansion of the RF in the temporal dimension through the convolutional layers, without changing the depth of the network. We express the size of the convolutional kernels in the temporal (T) and spatial (S^2) dimensions as $T \times S^2$. The 7×7^2 convolutional kernel in the first layer is replaced with a convolutional kernel of size 3×7^2 (1×7^2 for Video BagNet-1). In the following layers, we modify a set of 3D ResNet-50 bottleneck blocks. Bottle-

neck blocks consist of three consecutive convolutional layers of size $\begin{bmatrix} 1 \times 1^2, \\ 3 \times 3^2, \\ 1 \times 1^2 \end{bmatrix}$. We

replace them with $\begin{bmatrix} 1 \times 1^2, \\ 1 \times 3^2, \\ 1 \times 1^2 \end{bmatrix}$.

In addition, to prevent the temporal RF size from growing in the first layer, we alter the MaxPool operator that follows layer *conv1* to perform pooling only in the

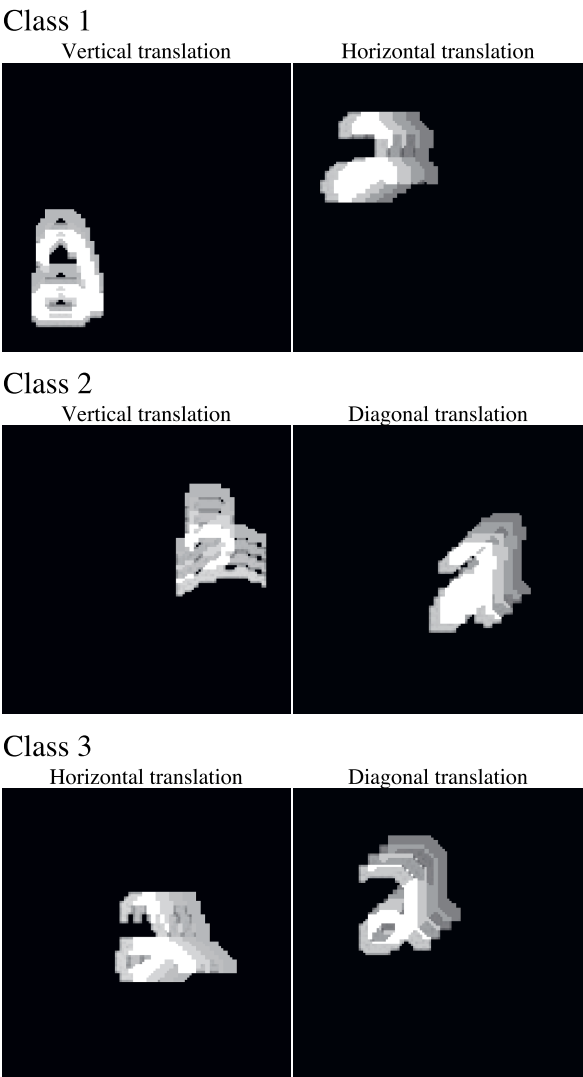


Figure 6.2: Example of videos of digit 2 from the *Directional Moving MNIST* dataset. The videos are composed of two sub-actions, i.e. vertical, horizontal or diagonal translation. Sub-action co-occurrences determine the video class. We explicitly superimposed multiple frames with shading to show the movement.

spatial dimensions. To maintain a comparable amount of parameters between 3D ResNet-50 and the different Video BagNet models, we widen the number of channels. Finally, to keep the input size equal to the video length, we remove the padding. Table 6.1 provides an overview of the architecture design of Video BagNet vs. 3D ResNet-50.

6.4 EXPERIMENTS

6.4.1 DATASETS

We study the effect of the temporal RF size on two long-term datasets, namely the *Directional Moving MNIST*, that we propose, and MultiTHUMOS [1]. These datasets contain multiple sub-actions and can last up to several minutes. For these datasets, the classification task consists of recognizing the sub-actions that compose the videos.

Directional Moving MNIST is a dataset composed of videos of one single moving digit, randomly sampled from the original MNIST dataset [30]. It contains 3 classes and 1000 videos per class. In this dataset, the digit translations correspond to sub-actions and the co-occurrence of two sub-actions determines the video class. More specifically, vertical and horizontal translation form class 1, vertical and diagonal translation form class 2 and horizontal and diagonal translation form class 3. Within each class, digit appearance and starting position have been randomized. In addition, the translations occur at two possible speeds. All sub-actions have equal duration and there are no pauses between consecutive sub-actions.

One fixed sub-action order appears in the training set. At test time we use two sets: in the *test set without permutations*, the sub-action order is the same as training time; while in the *test set with permutations* the sub-action order is permuted with 50% probability. An example of the *Directional Moving MNIST* dataset is provided in Figure 6.2.

MultiTHUMOS [1] is a multi-label video dataset for long-term action recognition. It is a collection of 400 complex, unconstrained, sports videos that have been densely annotated with sub-action time steps. The dataset contains a total of 65 possible sub-actions and each video contains, on average, 84.03 ± 113.56 sub-actions. The small size of the dataset prevents from training classification models using all the possible sub-action combinations and orders that usually occur in sports videos. For example, the dataset contains 20 basketball videos of which 15 videos contain the sub-actions *BasketballDribble*, *Run*, *BasketballPass*. Only 4 videos contain the order *BasketballDribble - Run - BasketballPass*.

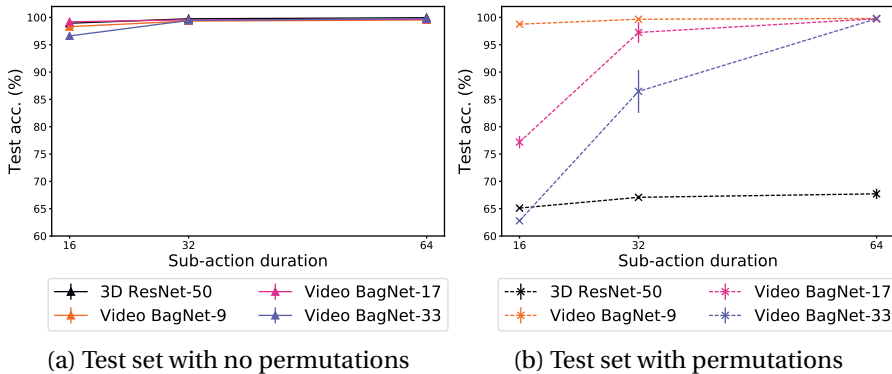


Figure 6.3: Sensitivity to sub-action order on the *Directional Moving MNIST* dataset. Models with different temporal RF are tested on two test sets with the same order (a) and different order (b) w.r.t. training time. The models with small temporal RF compared to the sub-action duration, namely Video BagNet 9, 17 and 33, perform well on the two sets. Differently, 3D ResNet, with temporal RF larger than 100 frames, overfits the temporal order at training time and fails to classify the test set with permutations.

6.4.2 THE SIZE OF THE TEMPORAL RF AFFECTS MODEL SENSITIVITY TO SUB-ACTION ORDER

We design a simple controlled experiment to investigate whether spatio-temporal models encode the sub-action order through their temporal RF. For this, we deploy the *Directional Moving MNIST* dataset. We vary the size of sub-actions to relate it with different temporal RF sizes. Specifically, we use sub-action duration of 16, 32 or 64 frames and temporal RF size equal to 217 frames for 3D ResNet-50 and 9, 17 and 33 frames for our Video BagNet.

The results of this experiment are summarized in Figure 6.3. Irrespectively of the temporal RF size and the sub-action duration, all the models perform well when the order of sub-actions of the training and test sets match, that is in the *test set without permutations*. However, on the *test set with permutations*, the models with large temporal RF size compared to the sub-action duration, e.g. 3D ResNet-50, and, in some instances, Video BagNet-17 and Video BagNet-33, perform poorly. In particular, 3D ResNet-50 always achieves an accuracy of $\sim 66\%$, which is equivalent to classifying correctly the videos with no permutations ($\sim 50\%$ of the *test set with permutations*) and randomly the videos with sub-action permutations. Our

Video BagNet-9, which has the shortest temporal RF among the analyzed models, performs above 98.5% on all the different test videos.

These results show that sensitivity to sub-action order depends on the sub-action duration and temporal RF size. We quantify the sensitivity to order by relating the sub-action size to the temporal RF size. For this, we analyze the convolutional shifting windows in the last convolutional layer of the 3D ResNet-50 and Video BagNet models, represented in Figure 6.1. In particular, we measure the sensitivity by a ratio of the amount of shifting windows that contain single sub-actions (*# single sub-action windows*) over the total amount of convolutional windows (*# total windows*). When the ratio is high, the sensitivity to the sub-action order is low. As shown in Figure 6.1, models with very large temporal RF size, like 3D ResNet-50, always see sub-action co-occurrences rather than single sub-actions. Therefore, in Figure 6.4, their ratio *# single sub-action windows / # total windows* is always low, which leads to low performance on the test sets with permutations. On the other hand, models with small temporal RF size, e.g. Video BagNet-9, have a large ratio of *# single sub-action windows / # total windows* and low sensitivity to the sub-action order, achieving good performance on the test set with permutations.

6.4.3 SMALL VS. LARGE TEMPORAL RF FOR LONG-TERM VIDEO ACTION RECOGNITION

6

Model	Temporal RF	mAP
Single-frame CNN [31]	1	25.4
MultiLSTM [1]	15	29.7
3D ResNet-50 [4]	>100	22.45
Video BagNet-33	33	26.37
Video BagNet-17	17	28.97
Video BagNet-9	9	30.21
Video BagNet-1	1	12.60

Table 6.2: Classification accuracies of models with small and large temporal RF on the MultiTHUMOS dataset. We compared our evaluated models (bottom rows) to the baselines proposed in [1] (top rows). Despite being trained from scratch, our Video BagNet models with temporal RF 9, 17 and 33 perform comparably to the ImageNet [32] pre-trained baselines. Models with smaller temporal RF, e.g. Video BagNet-9, recognize sub-action occurrences and ignore temporal order, achieving the best performance. Video BagNet-1 cannot model motion by seeing just single frames, which has the lowest mean average precision.

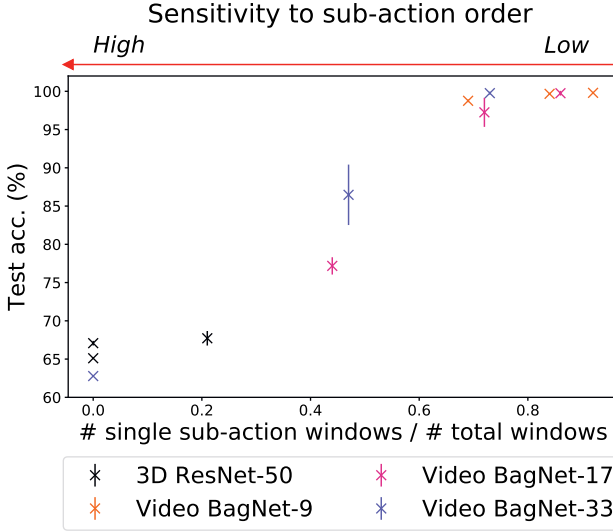


Figure 6.4: Accuracy on the *Directional Moving MNIST* test set with permutations in terms of models sensitivity to sub-action order. Sensitivity to sub-action order depends on the sub-action duration and temporal RF size, as shown in Figure 6.1. It can be expressed by counting the amount of convolutional shifting windows that contain single sub-actions (*# single sub-action windows*) over the total convolutional windows (*# total windows*). Models with large ratio *# single sub-action windows* / *# total windows*, like Video BagNet-9, are less sensitive to order and achieve good performance. Models with very large temporal RF sizes, like 3D ResNet-50, always see sub-action co-occurrences rather than single sub-actions. Therefore, their ratio *# single sub-action windows* / *# total windows* is low and their order sensitivity is high, thus performing poorly on the test set with permutations.

In our controlled experiment, we show that models with large temporal RF encode the sub-action order at training time. We argue that this causes misclassification when the distributions of sub-actions order are different in the training and test sets. This is the case for the commonly used MultiTHUMOS dataset, which only consists of 400 videos with high variability in sub-actions composition and order.

We evaluate the effect of the temporal RF size on MultiTHUMOS. Again, we deploy 3D ResNet-50 and Video BagNet with temporal RF 1, 9, 17 and 33. We train the models from scratch, without using either pre-training or data augmentation. We train with 512 input frames, with batch size 4. We do this to limit the com-

putational effort of our experiments. Since we train the models from scratch and without data augmentation, our results are not comparable to current state-of-the-art [33]. Nevertheless, employing this fixed experimental setup for all the analyzed models allows us to fairly compare different temporal RF sizes.

The results in Table 6.2 show that models with small temporal RF size outperform models with large temporal RF size on this dataset. The highest accuracy is obtained with Video BagNet-9. These results suggest that encoding long-term information, including sub-action order, is hurting the classification of MultiTHUMOS. This long-term information could correspond to the precise order of sub-actions or to the varying durations of different sub-actions. This is sensible: the multi-label classification problem of MultiTHUMOS consists in recognizing all the single sub-actions occurring in a video. Sub-action classification can be achieved by looking at short temporal extents that contain the sub-action. Because of the high variation in the temporal composition of sports videos, overemphasizing long-term information is not necessary or even decreases the sub-action recognition accuracy. On the other hand, for Video BagNet-1 it shows that if the model encodes neither long-term nor short-term information, the accuracy decreases. The results indicate that the short-term information captured by small temporal RF seems essential for good classification performance.

We find that our results are comparable to the baseline models proposed in [1], as illustrated in Table 6.2. It is worth noting that the single-frame CNN [31], which cannot model temporal information by design, has the advantage of being pre-trained on ImageNet [32], thus explaining the superior performance compared to Video BagNet-1. Similarly, the MultiLSTM model [31] uses pre-trained image features. Despite the lack of pre-training, Video BagNet-9 and 17 achieve 28.97% and 30.21% mAP, which is similar to mAP of 29.7% mAP obtained by Video MultiLSTM.

6.5 CONCLUSIONS

In this paper, we investigate whether spatio-temporal models for long-term action recognition encode sub-action order through their temporal RF. Our experiments reveal that when the temporal RF size is larger than the sub-action duration, the models are sensitive to the sub-action order. We provide a measure for the sensitivity to the sub-action order by a ratio of the number of convolutional windows that contain single sub-actions over the total number of convolutional windows. A higher ratio makes the models less sensitive to the sub-action order.

Sensitivity to sub-action order causes misclassification when the order of sub-actions are different during training and test time. This might occur in long-term action recognition, since it is difficult to collect training samples containing all the

sub-action permutations that exist in natural videos. We show that small temporal RFs are robust to permutations of sub-actions, which is beneficial when limited sub-action orders are available at training time. Our study is conducted on 3D convolutional networks. Nevertheless, the conclusions could be generalizable to other spatio-temporal models that use the RF to encode temporal dependencies.

Acknowledgements. This work is part of the research program Efficient Deep Learning (EDL), which is (partly) financed by the Dutch Research Council (NWO).

REFERENCES

- [1] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. “Every moment counts: Dense detailed labeling of actions in complex videos”. In: *International Journal of Computer Vision* 126 (2018), pp. 375–389.
- [2] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. “Hollywood in homes: Crowdsourcing data collection for activity understanding”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 510–526.
- [3] W. Brendel and M. Bethge. “Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet”. In: *International Conference on Learning Representations*. 2018.
- [4] K. Hara, H. Kataoka, and Y. Satoh. “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 6546–6555.
- [5] N. Hussein, E. Gavves, and A. W. Smeulders. “Timeception for complex action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 254–263.
- [6] N. Hussein, M. Jain, and B. E. Bejnordi. “Timegate: Conditional gating of segments in long-range activities”. In: *arXiv preprint arXiv:2004.01808* (2020).
- [7] X. Liu, S. L. Pinteá, F. K. Nejadasl, O. Booi, and J. C. van Gemert. “No frame left behind: Full Video Action Recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14892–14901.
- [8] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei. “Learning spatio-temporal representation with local and global diffusion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12056–12065.
- [9] X. Wang, R. Girshick, A. Gupta, and K. He. “Non-local neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7794–7803.
- [10] C. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krähenbühl, and R. B. Girshick. “Long-Term Feature Banks for Detailed Video Understanding”. In: *CoRR* abs/1812.05038 (2018). arXiv: 1812.05038. URL: <http://arxiv.org/abs/1812.05038>.
- [11] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. “Temporal relational reasoning in videos”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 803–818.

- [12] J. Carreira and A. Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset”. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.
- [13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. “Learning spatiotemporal features with 3d convolutional networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4489–4497.
- [14] N. Hussein, E. Gavves, and A. W. Smeulders. “Videograph: Recognizing minutes-long human activities in videos”. In: *arXiv preprint arXiv:1905.05143* (2019).
- [15] G. Bertasius, H. Wang, and L. Torresani. “Is space-time attention all you need for video understanding?” In: *ICML*. Vol. 2. 3. 2021, p. 4.
- [16] M. Patrick, D. Campbell, Y. M. Asano, I. M. F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques. *Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers*. 2021. arXiv: 2106.05392 [cs.CV].
- [17] G. Sharir, A. Noy, and L. Zelnik-Manor. “An Image is Worth 16x16 Words, What is a Video Worth?” In: *CoRR* abs/2103.13915 (2021). arXiv: 2103.13915. URL: <https://arxiv.org/abs/2103.13915>.
- [18] N. Hussein, E. Gavves, and A. W. Smeulders. “Pic: Permutation invariant convolution for recognizing long-range activities”. In: *arXiv preprint arXiv:2003.08275* (2020).
- [19] K. Hara, H. Kataoka, and Y. Satoh. “Learning spatio-temporal features with 3d residual networks for action recognition”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 3154–3160.
- [20] H. Kuehne, A. Arslan, and T. Serre. “The language of actions: Recovering the syntax and semantics of goal-directed human activities”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 780–787.
- [21] O. Strafforello, K. Schutte, and J. van Gemert. “Are current long-term video understanding datasets long-term?” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2023.
- [22] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. “Actionvlad: Learning spatio-temporal aggregation for action classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 971–980.
- [23] D.-A. Huang, V. Ramanathan, D. Mahajan, L. Torresani, M. Paluri, L. Fei-Fei, and J. C. Niebles. “What makes a video a video: Analyzing temporal information in video understanding models and datasets”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7366–7375.
- [24] G. A. Sigurdsson, O. Russakovsky, and A. Gupta. “What actions are needed for understanding human actions in videos?” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2137–2146.
- [25] T. Shen, Y. Huang, and Z. Tong. “FaceBagNet: Bag-Of-Local-Features Model for Multi-Modal Face Anti-Spoofing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019.

- [26] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord. “Learning representations by predicting bags of visual words”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6928–6938.
- [27] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [28] S. Ioffe and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.
- [29] V. Nair and G. E. Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *Icml*. 2010.
- [30] L. Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.
- [31] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. “ImageNet Large Scale Visual Recognition Challenge”. In: *IJCV* 115 (2015), pp. 211–252.
- [33] J. Zhou, K.-Y. Lin, H. Li, and W.-S. Zheng. “Graph-based high-order relation modeling for long-term action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8984–8993.

7

DISCUSSION

Goodhart’s Law: “*When a measure becomes a target, it ceases to be a good measure.*”

In this thesis, we investigate whether the development of computer vision solutions is always in line with the intended objectives set by humans. We do this with a focus on two particular applications, namely object detectors meant as user assistive tools and long-term action recognition in videos. This research reveals cases when standardized training and evaluation methods are sub-optimal with respect to the end use of the computer vision algorithms. In addition, we uncover shortcut learning potential in current video datasets. Here we outline our main conclusions and their implications.

7.1 OBJECT DETECTORS FOR USER ASSISTANCE

IoU and AP are insufficient to evaluate object detectors In many applications, including industrial visual inspection, anomaly detection, or medical imaging, human tasks are facilitated with automatic object detections. For this utilization of object detectors, the quality of the object detectors output should be perceived as high by human users. In Chapters 2 and 3, we evaluate if the predicted *object location* by object detectors aligns with what humans consider a well detected object. In Chapter 2, we performed a fully controlled experiment, where we asked humans to evaluate pairs of object detections with the same localization error, measured with the IoU metric, but different in size or location. Our results showed that humans prefer larger over smaller object detections with the same location error. In addition, for asymmetric objects the position of the detection matters, since it helps humans better identify the object. However, the human preference that we discovered is not captured by the IoU metric, which is to some degree insensitive to spatial translations and size differences of the detections. The same conclusion applies to the AP metric, which relies on the IoU as a measure for the localization

accuracy. In Chapter 3, we confirmed that also real, commonly used, object detectors predict large and small detections equally often, therefore they are also not aligned with human preference. We find that training with our proposed asymmetric loss, or upscaling the detections in a post-processing step, results in higher user preference [1]. Finally, humans consistently prefer larger object detections, even when the AP approximately goes to zero. This suggests that, in our confined experimental setup, the AP does not correlate with human quality judgments of object detectors. Although we lack sufficient evidence to generalize this conclusion, there may be other scenarios where AP and user preferences do not align. We urge the future development of object detectors to assist humans to consider incorporating qualitative assessments, to ensure a more comprehensive understanding of the model performance, and explore alternative training and evaluation metrics. The results obtained in Chapter 2 and 3 suggest that standard object detector development methods do not serve well the objective of developing valid user assistive tools. In fact, training and evaluating based on the common IoU and AP metrics might result in sub-optimal detections with respect to human perception of detection quality.

Beyond controlled environments and assistive object detectors The work described in Chapters 2 and 3 is done in a confined experimental environment, restricted to the assessment of single objects in MS COCO images [2]. Future work should explore how object detectors are perceived when they are employed to assist users in a real-world application, for example in the medical and industrial domains. Similarly to the discrepancies that we observed with IoU and AP [1, 3], we wonder if other mismatches between human assessments and quantitative evaluations exist in other applications of computer vision, such as video editing tools [4], object tracking and human activities recognition algorithms in socially assistive robots [5], and behavioral imaging, a technique used to monitor and diagnose behavioral disorders [6, 7]. We encourage future research to investigate this direction further.

The impact on user trust Especially in medical applications, it is paramount that the outcomes of the algorithm match high user quality perception to promote trust in the assistive system. Our findings in Chapter 2 and 3 indicate that standardized computer vision practices might be sub-optimal for this purpose. Recent work shows that different type of assistive object detectors errors have different effects on user trust [8]. For example, precision and recall errors in object detections have a larger impact on trust compared to localization errors. In addition, Barbosa *et al.* [9] point out that both the evaluation metrics and the way they are presented

to users affect the trust in computer vision models. Future research is essential to explore whether aligning the object detections with human quality judgments, by training with our asymmetric loss function or simply by upscaling the predicted boxes, also increases the trust of users in the assistive object detection systems.

7.2 LONG-TERM ACTION RECOGNITION

Spatial information is more important than temporal information in long-term behavior recognition In Chapter 4, we investigated whether including spatial and temporal attention mechanisms enhances the results of long-term human behavior recognition in videos. We found that adding a spatial attention module on top of a 3D convolutional backbone improves the recognition performance. Furthermore, we observe that the spatial region centered around the human subject contains the most discriminative information. Indeed, allowing the model to explicitly focus on this 'actor-focused' region further enhances the results. On the other hand, the addition of temporal attention does not significantly impact the model performance. We find this counter-intuitive, as enhancing the temporal modeling capabilities should help encoding the complex temporal patterns ongoing in long-term videos.

Encoding long-term information is not necessary to solve current video datasets

In a soccer game, following the players actions over time and counting the goals is required to track which team is winning. Similarly, long-term temporal information is essential to understand long-term human activities, like the progress of a medical surgery or the assessment of behaviors recorded by surveillance cameras. However, the results from Chapter 4 seem to indicate that temporal information is superfluous in long-term action recognition models. To further investigate this phenomenon, in Chapter 5, we proposed a method to evaluate whether long-term temporal information is needed to classify long-term action videos. We apply this method to three commonly used datasets, namely Breakfast [10], CrossTask [11] and LVU [12]. Surprisingly, we found that long-term information is not necessary for the majority of the analyzed videos. In fact, looking solely at a single short video segment is sufficient to correctly classify the long-term actions. This conclusion suggests that modeling long-term information, including, for example, the order and duration of the short-term actions in the videos, is not necessary. In Chapter 6, we showed that 3D convolutional neural networks with large temporal receptive fields can encode temporal order information and overfit to specific orders when small datasets are available at training time. These models performed worse than 3D convolutional neural networks with a *small* temporal receptive on

the MultiTHUMOS dataset [13]. Also in this case, encoding long-term temporal information does not enhance, and even deteriorates, the long-term action recognition performance.

Long-term action datasets should require long-term reasoning Our results from Chapter 5 and 6 showed that current long-term action video datasets do not require long-term temporal information to be solved. Thus, computer vision models trained and evaluated on these datasets are not exploiting long-term temporal dynamics, but pick short-term shortcuts while showing promising accuracy scores. On the other hand, long-term information is fundamental to understanding different human behaviors, like in the aforementioned examples related to sports, health and surveillance. Models that are not capable of encoding long-term temporal dynamics are unlikely to perform successfully on the recognition of these types of long-term actions. We urge computer vision researchers to collect and use more representative video datasets to investigate the challenging problem of long-term action recognition. A good example of long-term action videos could be amateur sport videos, where the class to predict is the team that is winning. A different approach to encourage long-term reasoning and refrain from shortcut learning could involve transitioning from video classification to tasks that demand a more nuanced understanding of temporal dynamics. Examples of these tasks are video captioning [14], where models are trained to generate a textual summary of the video content, or video question answering [15], where models are trained to answer questions pertaining to the video content. By asking questions related to different moments in the videos, the models would be forced to encode temporal information over a long time span.

7.3 TRAINING AND EVALUATING COMPUTER VISION MODELS

Solving a datasets is not equal to solving a computer vision task The last years have witnessed an explosion of models that compete to achieve the best accuracy scores on common computer vision benchmarks, like ImageNet [16]. As discussed in Chapter 1, solving ImageNet does not necessarily correspond to solving the broader task of automatic image classification. Similarly, the optimization of object detection algorithms through conventional training and evaluation metrics, on datasets like MS COCO [2], does not guarantee obtaining effective general-purpose object detectors. For example, the IoU and AP metrics do not accurately capture human preferences (Chapters 2 and 3). Along the same vein,

currently used datasets in long-term action recognition research do not promote long-term reasoning in computer vision models, but rather the use of unintended short-term shortcuts (Chapters 5 and 6). While the developed models show high accuracy scores when tested on standard benchmarks, they fail to generalize in real-world scenarios, where long-term reasoning is crucial. We saw that the common training and evaluation metrics or datasets might not provide an extensive assessment of the performance of computer vision algorithms. A more in-depth analysis of a model performance should consider additional criteria, especially an interpretability analysis to understand *why* an algorithm is making a certain prediction, even when the prediction seems accurate. This step is crucial to understand if a computer vision model is looking at the intended features to make a prediction, or exploiting unexpected shortcuts. Visual explanation tools, like Grad-CAM [17], highlight the input regions that lead a model to a specific prediction. These methods could reveal whether a seemingly accurate long-term action recognition model is exploiting temporal information or not. Similar interpretability tools should become a standard practice in the evaluation of computer vision models. Furthermore, evaluating a computer vision algorithm with respect to its intended application is essential. For tasks that involve humans as the end-users, conventional quantitative evaluation metrics might be insufficient. In these cases, a qualitative evaluation of the results of computer vision should be included.

7.4 FINAL WORDS

Progress in computer vision has shown revolutionizing applications in fields such as healthcare [18–20], autonomous driving [21, 22] and environmental monitoring [23–25]. Not only can computer vision techniques streamline laborious human tasks, but they can also have significant social and environmental impacts, such as facilitating early medical diagnoses, improving traffic safety, and safeguarding marine ecosystems. However, for the successful integration of computer vision into human workflows, and even in daily lives, establishing user trust in computer vision systems is paramount. For this, we need to be sure that the models are developed securely and effectively. In this dissertation, we demonstrated that the current paradigm for developing computer vision solutions, and for the task that they are being deployed for, may not always yield the intended results. We advocate future research towards an approach that goes beyond merely striving for common accuracy metrics on standard benchmark datasets.

REFERENCES

- [1] O. Strafforello, O. S. Kayhan, O. Inel, K. Schutte, and J. C. van Gemert. “Aligning object detector bounding boxes with human preference”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2024.
- [2] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft coco: Common objects in context”. In: *ECCV*. Springer, 2014.
- [3] O. Strafforello, V. Rajasekar, O. S. Kayhan, O. Inel, and J. van Gemert. “Humans disagree with the IoU for measuring object detector localization error”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1261–1265.
- [4] A. Duico, O. Strafforello, and J. van Gemert. “Can we predict the Most Replayed data of video streaming platforms?” In: *arXiv preprint arXiv:2309.06102* (2023).
- [5] V. M. Montaña-Serrano, J. M. Jacinto-Villegas, A. H. Vilchis-González, and O. Portillo-Rodríguez. “Artificial Vision Algorithms for Socially Assistive Robot Applications: A Review of the Literature”. In: *Sensors* 21.17 (2021), p. 5728.
- [6] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin. “Behavioral imaging and autism”. In: *IEEE Pervasive Computing* 13.2 (2014), pp. 84–87.
- [7] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. M. Farinella. “Computer vision for assistive technologies”. In: *Computer Vision and Image Understanding* 154 (2017), pp. 1–15.
- [8] S. de Witte, O. Strafforello, and J. van Gemert. “Do Object Detection Localization Errors Affect Human Performance and Trust?” In: *arXiv preprint arXiv:2401.17821* (2024).
- [9] G. D. J. Barbosa, D. dos Santos Ribeiro, M. do Carmo Silva, H. Lopes, and S. D. J. Barbosa. “Investigating the relationships between class probabilities and users’ appropriate trust in computer vision classifications of ambiguous images”. In: *Journal of Computer Languages* 72 (2022), p. 101149.
- [10] H. Kuehne, A. Arslan, and T. Serre. “The language of actions: Recovering the syntax and semantics of goal-directed human activities”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 780–787.
- [11] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic. “Cross-task weakly supervised learning from instructional videos”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3537–3545.

- [12] C.-Y. Wu and P. Krahenbuhl. "Towards long-form video understanding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1884–1894.
- [13] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. "Every moment counts: Dense detailed labeling of actions in complex videos". In: *International Journal of Computer Vision* 126 (2018), pp. 375–389.
- [14] V. Jain, F. Al-Turjman, G. Chaudhary, D. Nayar, V. Gupta, and A. Kumar. "Video captioning: a review of theory, techniques and practices". In: *Multimedia Tools and Applications* 81.25 (2022), pp. 35619–35653.
- [15] L. Zhu and Y. Yang. "Actbert: Learning global-local video-text representations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8746–8755.
- [16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.
- [18] B. Hunter, S. Hindocha, and R. W. Lee. "The role of artificial intelligence in early cancer diagnosis". In: *Cancers* 14.6 (2022), p. 1524.
- [19] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher. "Deep learning-enabled medical computer vision". In: *NPJ digital medicine* 4.1 (2021), p. 5.
- [20] A. Sharma, K. Prasad, S. V. Chakrasali, D. Gowda, C. Kumar, A. Chaturvedi, and A. A. J. Pazhani. "Computer vision based healthcare system for identification of diabetes & its types using AI". In: *Measurement: Sensors* 27 (2023), p. 100751.
- [21] J. Janai, F. Güney, A. Behl, A. Geiger, *et al.* "Computer vision for autonomous vehicles: Problems, datasets and state of the art". In: *Foundations and Trends® in Computer Graphics and Vision* 12.1–3 (2020), pp. 1–308.
- [22] H. Gajjar, S. Sanyal, and M. Shah. "A comprehensive study on lane detecting autonomous car using computer vision". In: *Expert Systems with Applications* 233 (2023), p. 120929.
- [23] E. V. Sheehan, D. Bridger, S. J. Nancollas, and S. J. Pittman. "PelagiCam: A novel underwater imaging system with computer vision for semi-automated monitoring of mobile marine fauna at offshore structures". In: *Environmental monitoring and assessment* 192 (2020), pp. 1–13.
- [24] C. Xia, L. Fu, Z. Liu, H. Liu, L. Chen, and Y. Liu. "Aquatic toxic analysis by monitoring fish behavior using computer vision: A recent progress". In: *Journal of toxicology* 2018 (2018).

- [25] B. Koger, A. Deshpande, J. T. Kerby, J. M. Graving, B. R. Costelloe, and I. D. Couzin. “Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision”. In: *Journal of Animal Ecology* (2023).

ACKNOWLEDGMENTS

During the years of my PhD I had the privilege of meeting, interacting and becoming friends with people not only exceptionally smart, but also humble, easy going, and kind. I will forever be grateful for this.

To my supervisors, **Jan** and **Klamer**, thank you for believing in my capabilities and continuously motivating me, even during the difficult COVID times. I am deeply thankful that I got to have such knowledgeable and thoughtful supervisors. Jan, I now find myself asking “Why?” and “What’s the storyline?” whenever I approach research. Klamer, thank you for teaching me to always be critical and question my work and for always finding time for our Friday meetings.

Marcel, thank you for your valuable suggestions and for your timely feedback throughout these 4+ years. I deeply appreciate how you always find time to be present when discussing important matters in the PRB group, despite your busy schedule.

Dear paranymphs, **Attila** and **Robert-Jan**, I am so grateful that our trip to Tübingen marked the beginning of our friendship. It has been so nice and fun to share the office with you, all the borrels, the ski-trip, the surfing in Porto the trips to London and Paris. Attila, thank you for being the initiator of the PRB ski-trip, for being a patient surf instructor, for all the fun things you organized outside the office and for involving me in your brilliant NeurIPS paper. Your energy and friendliness are unique. Robert-Jan, thank you for always trying to make the PRB group cohesive, making everyone feeling comfortable and involved, and always bringing board games. Your thoughtfulness is inspiring. I am honored I attended your and **Fleur**’s wedding and I wish you a lifetime of happiness. And to both of you, Attila and Robert-Jan, thank you for accepting me despite the thousands of cluster-related questions! I will miss you greatly, but I know we will stay in touch.

Oana and **Osman**, it has been a great pleasure collaborating with you. Oana, thank you for sharing with me your valuable knowledge about crowdsourcing, since my master thesis. Osman, thank you for the coffee breaks, your visits in the office, the fun time at ICCV 2023 and all the inspiring conversations. I am grateful that you presented our paper at ACVR this year. **Marcos**, discussing video understanding with you in the first months of my PhD was enlightening. **Petr**, collaborating with you and talking about temporal receptive fields was refreshing.

Xin, it was comforting to have a colleague who worked on my same topic and I

am happy that we wrote a paper together. It was nice that you joined the spring school in Guimarães. **Casper**, I am happy that we attended the spring school together and shared great times in Portugal, London, and Paris. Thank you for your energetic vibes, and for being such a talented bouldering instructor. **Seyran**, I am grateful that we collaborated on the Intelligent Promo Generation project, I learnt a lot from your supervision. I find your current research so fascinating and I am excited to see all the results it will bring. Also, thank you for organising such a delicious Persian dinner experience in Paris.

Silvia and **Nergis**, thank you for being so talented and yet so kind and easy going. Silvia, your cat is precious. Nergis, I am glad we shared the office in the first months of my PhD, that I met adorable baby Volkan and attended his first birthday party. **Hadi**, thank you for all the bright suggestions you shared in the CV lab meetings. It was nice to attend ICCV 2023 together. **Xucong**, I appreciate your down to earth attitude and your humor. I am glad you joined the CV lab during my PhD.

Sander, thank you for entertaining me with adorable cat pictures, for your company in the office and during ICCV in Paris. **Amogh** and **Marian**, thank you for sharing your PhD journey stories with me and the PRB group. It was fun to organise the retreat talk with you. Marian, thanks for being an active member of the Writing Workshop and for your shared passion for cats. Amogh, I hope I will see you in more festivals. **Yancong**, you were an inspiring senior PhD student in the first years of my PhD. How great that you joined the first PRB sky-trip! **Ziqi**, it was nice to show each other our cats' pictures and to meet you in London. **Yunqiang**, it was a pleasant surprise to meet you on the train from Leuven and to have a chance to catch up. **Yeshwanth**, when you supervised my Deep Learning project in my first year of master I could not have imagined that we would have become colleagues!

Aurora and **Alejandro**, I am glad we became friends outside the office! Thank you for joining the ski-trip and all the bouldering sessions. Alejandro, thank you for bringing gym motivation and for sending regularly amazing drawing videos and mind-blowing astrophysics talks. It is nice to share so many hobbies. Aurora, in you I found my dancing buddy. Thank you for visiting me in Italy and for cat sitting my boys. I am confident we will be friends for a long time to come.

Xiangwei, thank you for introducing me to sweet Chinese bao and for managing the calendar so well. **Hesam**, thank you for joining the bouldering sessions and for the interesting presentations on event videos. **Chengming**, you were a fun presence at NCCV 2024. **Thomas**, **Sayak**, even though we did not share so much time together as members of the group, I think that you have been a very nice addition to the CV lab. **Akshit**, nice that you started joining the CV lab meetings.

Jordan, it was nice to occasionally have you in the office and I wish you luck in the rest of your studies. Also, thank you for taking us to such a fun techno party in

Paris. **Fatemeh**, it was such a nice experience to go to ICCV in Paris together! You are such a fun and talented person.

David and **Marco**, thank you for inspiring the PRB group to being supportive and informal. David, thank you for always giving me valuable feedback in my presentations and for your unfailing presence at the lunch table. Marco, it took me a while to understand your sarcasm, but when I did I found it an amazing quality. **Jesse**, thank you for your feedback in the lab meetings, for giving interesting talks and for your variable insights in my poster presentation. **Hayley**, your research is so inspiring to me. I am grateful for your presence in the lab: you always gave me precious feedback and helped me discover my research interests. **Tom**, I wish I were one of the students who joined the field trip to Copenhagen! Thank you for all the fun times and for introducing me to techno parties. We do need to plan that Berlin trip. **Taylan**, thank you for taking me to the drum and bass party and for impressing me with your natural bouldering talent.

Rickard, I wish you had started your PhD before me, so I could have learnt from you everything about how to organise the perfect research visit and give very entertaining talks. I am glad we became friends, that we already went to two ski-trips together, and that **Mirthe** joined, too. I hope this will become a long-lasting tradition. Also, thank you for taking me to techno festivals and sharing music recommendations! **Jim** and **Myrthe**, it was nice to count on your presence in the office and at the borrel on Thursdays. I enjoyed Fiesta Macumba with you! **Ramin** and **Mahdi**, talking to you in the office was always nice. Ramin, you are an inspirational job hunter. Mahdi, it was so kind of you to gift me Persian saffron and sweets. **Gijs**, thank you for your amazing cooking at the PRB Retreat 2024!

Vandana, **Arman**, **Mo**, thank you for welcoming me in your office in my occasional visits this year. Arman, how cool that we both did our master thesis at IBM CAS and our PhD in the PRB group. **Chirag**, thank you for making sure that all the newcomers are always properly introduced! Also thanks for all you valuable advice and for trying your best to master the Italian accent. **Bernd** and **Tiffany**, your knowledge on the intersection between AI and humans has been inspiring for me. **Jose**, so cool that we are defending our PhD in the same month! **Chenxu** and **Zhi-Yi**, thanks for organising a successful PRCVSPC retreat! **Era**, it was nice to meet you during the last retreat. **Merve** and **Jing**, I am glad I got to meet you before finishing my PhD. Merve, I admire your sense of style. Jing, thank you for being an active member of the Thursday borrels.

Swier and **Gabriel**, I really appreciated going to the ski-trips with you and getting to know you better. Swier, thank you for being a patient snowboard instructor, but also a tenacious Murder Game player. Gabriel, thank you for introducing us to your amazing Käsespätzle. To **Stavros**, **Amelia**, **Yasin**, **Aysun**, **Ramin**, **Paul**, **Colm** and all **the Bioinformaticians**, it was always nice to chat with you at the coffee

machine and at the Thursday borrels. **Sara**, I had so much fun at the PRB board game session I attended, thanks for organising it!

Ruud and **Bart**, thank you for always be nice and helpful every time I knocked on your door with technical problems. **Saskia** and **Marunka**, thank you for your assistance throughout these years, despite my countless emails. **Azza**, your efforts in the cluster management is precious to us PhD students.

To my MSc students **Debadeep**, **Godwin**, **Alessandro**, **Sven**, **Marijn** and all the **BSc students** that I co-supervised at TU Delft, it was a real pleasure to see you successfully defending your theses. Thank you for your hard work.

To **TNO's Intelligent Imaging group**, thank you for making me feel welcome every time I was in the office. **Gertjan**, thank you for your positive energy and your excitement towards novel scientific research. **Sabina**, I am happy that we met at ICCV. **Luca** and **Thomas**, I am glad that I was involved in your thesis supervision. You really did great work.

Stefano and **Alberto**, we survived some tough time together. I know that it would have been even harder without your presence in the house, the long bike rides, the workouts in the grass next-door. Stefano, thank you for all your funny jokes and for sharing La Zanzara quotes. Alberto, you were the first person who inspired me to become vegetarian. Thank you for that!

I am grateful that many of the friends I met during my bachelor and master remained my good friends throughout my PhD and to this day. **Ioannis**, we finally see the end of our PhDs! **Alessandro**, it was an amazing surprise to bump into you at ICCV 2023. I hope it will happen again in another conference. **Bárbara**, thank you for all our conversations, the dinner parties, your support when I was applying for my postdoc. I am looking forward to hearing about your San Francisco adventure. **Nirmal**, you and **Meghdipa** are the best cat sitters ever. The boys were lucky to stay with you. **Alessandro**, I am so thankful I got to be your flatmate, even if it was for a short time: I had the best time. Thank you for sharing with me gym motivation and for organising amazing dinners and bike trips. I hope there will be more of that. **Adrien**, thank you for visiting in Ventimiglia and Rotterdam, and even letting Bobi and Sebi sleep with you! Even though I never reply on WhatsApp, I know I can count on our friendship. **Akshay**, I am so happy you went all the way to Stanford and returned to Delft. Your talent and humility are inspiring. **Camille**, it was so nice to see you in Ventimiglia this summer. I hope one day we will go to a surfing school together. **Miki**, **Pooja**, **Fabian**, **Fabian**, **Chriss**, **Eugenio**, it was always fun seeing you.

Kia, your way of speaking Italian piqued my curiosity since the first day I saw you in the gym. Thank you for being such a fun companion, I am so grateful I was your paranymph. **Chiara**, I am grateful for your friendship and so excited that we will defend our PhDs in the same month. Luckily, we came a long way since that

tiny student apartment in Aan't Verlaat. I wish you and **Tom** all the best. **Frederiek** and **Frida**, thanks for being our guides in Oslo, for the bouldering sessions and for cycling together up and down the Hoek van Holland hill for 67 times. I wouldn't have done something that crazy without you. I wish you happiness in Norway. **Federica**, I still can't believe that we went from sitting next to each other in high school, to being flatmates during our bachelor in Turin, to doing our masters in the Netherlands and to both being PhD students at TU Delft. I am so grateful for so many years of friendship.

Iris, Betty and **Oma Joke**, thank you for being my second family in the Netherlands. **Eric, Sara, Angiolina**, thank you for being my second family in Italy. To my American family, **Michael, Lisa, Michaela** and **Angelina**, experiencing living in the U.S. with you has helped me throughout my life abroad. I am glad we met in Avignon and I hope I will come visit you soon.

To my family, thank you for your support through the years. My parents **Gianni** and **Isabella**, thank you for visiting me in the Netherlands, for sending regular pictures of Filippo, and for being the official party sponsors! **Amalia** and **Francesco**, thank you for visiting me in Rotterdam! **Chica** and **Virginia**, I am very grateful that you are coming to the Netherlands to see my PhD defense. **Pepi**, thank you for always showing interest in my research and for sharing interesting articles, and to you and **Lili**, thank you for watching my defense online. **Biancamaria** and **Mimo**, thank you for teaching me the importance of education since a young age.

To my old friends, **Carola, Laura, Margherita, Riccardo, Pietro, Martina, Gior-gia**, even though we do not see each other very often, I know I can count on your company in every summer and Christmas holiday. Thank you for your friendship throughout spacetime.

Jonathan, you truly are my biggest fan. Thank you for being my language assistant in - I confess - every cover letter I ever had to write. More importantly, thank you for your constant support, which extends far beyond my PhD, and for lightening my days since 2014.

Last but not least, **Bobì** and **Sebastian**, thank you for being my adorable buddies. Your company at home is invaluable.

To all the people who crossed my path during the last 4+ years: Thank you! You contributed meaningfully to my journey.

*Ombretta
Rotterdam, September 2024*

CURRICULUM VITÆ

Ombretta STRAFFORELLO

27 June 1995 Born in Bordighera, Italy.

EDUCATION

- 2019–2023 **PhD Computer Science**
Delft University of Technology, Delft, Netherlands
Thesis: Rethinking the objectives of computer vision systems
Promoters: Dr. J.C. van Gemert
 Dr. K. Schutte
 Prof. dr. ir. M.J.T. Reinders
- 2017–2019 **Master of Science in Computer Science**
Delft University of Technology, Delft, Netherlands
- 2014–2017 **Bachelor of Science in Computer Engineering**
Politecnico di Torino, Torino, Italy
- 2009–2014 **Pre-university education**
Liceo Scientifico A. Aprosio, Ventimiglia, Italy (2009–2014)
Skyline High School, Ann Arbor, Michigan, USA (2012–2013)

EXPERIENCE

- 2023–Present **Postdoctoral Researcher at KU Leuven, Leuven, Belgium**
- 2019–2023 **PhD Researcher at TNO, Den Haag, Netherlands**
- 2018–2019 **Research Intern at IBM, Amsterdam, Netherlands**

LIST OF PUBLICATIONS

In this thesis

1. O. Strafforello, V. Rajasekar, O. S. Kayhan, O. Inel, and J. van Gemert. “Humans disagree with the IoU for measuring object detector localization error”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 1261–1265
2. O. Strafforello, O. S. Kayhan, O. Inel, K. Schutte, and J. C. van Gemert. “Aligning object detector bounding boxes with human preference”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2024
3. L. Ballan, O. Strafforello, and K. Schutte. “Long-term Behaviour Recognition in Videos with Actor-focused Region Attention.” In: *VISIGRAPP (5: VISAPP)*. 2021, pp. 362–369
4. O. Strafforello, K. Schutte, and J. van Gemert. “Are current long-term video understanding datasets long-term?”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2023
5. O. Strafforello, X. Liu, K. Schutte, and J. van Gemert. “Video BagNet: short temporal receptive fields increase robustness in long-term action recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2023

Other publications

1. O. Strafforello, D. Soydaner, M. Willems, A.-S. Maerten, and S. De Winter. “Have Large Vision-Language Models Mastered Art History?”. In: *arXiv preprint arXiv:2409.03521* (2024)
2. O. Strafforello, G. M. Odriozola, F. Behrad, L.-W. Chen, A.-S. Maerten, D. Soydaner, and J. Wagemans. “BackFlip: The Impact of Local and Global Data Augmentations on Artistic Image Aesthetic Assessment”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2024)
3. S. de Witte, O. Strafforello, and J. van Gemert. “Do Object Detection Localization Errors Affect Human Performance and Trust?”. In: *arXiv preprint arXiv:2401.17821* (2024)
4. A. Lengyel, O. Strafforello, R.-J. Bruintjes, A. Gielisse, and J. van Gemert. “Color Equivariant Convolutional Networks”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023

5. J. Warchocki, T. Oprescu, Y. Wang, A. Damacus, P. Misterka, R.-J. Bruintjes, A. Lengyel, O. Strafforello, and J. van Gemert. "Benchmarking Data Efficiency and Computational Efficiency of Temporal Action Localization Models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2023
6. A. Duico, O. Strafforello, and J. van Gemert. "Can we predict the Most Replayed data of video streaming platforms?" In: *arXiv preprint arXiv:2309.06102* (2023)
7. P. Byvshev, R. J. Bruintjes, X. Liu, O. Strafforello, J. C. van Gemert, P. Mettes and Y. Xiao. 2022. "The Density-Extent Map of Video Representation Learning". (11 pages), Manuscript submitted for publication.

