

Plan van Aanpak

Maxim van Duin

Faculteit Digitale Media en Creatieve Industrie

Hogeschool van Amsterdam

Amsterdam, Nederland

maxim.van.duin@hva.nl

Samenvatting—De industrie kamp met een groeiend personeelstekort, mede door een gebrek aan scholing en ervaring. Veel bedrijven investeren in online trainingen, maar het gebrek aan de mogelijkheid om vragen te stellen belemmert effectieve kennisoverdracht.

Een virtuele lesassistent kan dit probleem oplossen door vragen te beantwoorden op basis van de lesstof. Traditionele chatbots schieten hierin tekort, terwijl Retrieval-Augmented Generation (RAG), gecombineerd met een Large Language Model (LLM), relevantere en beter onderbouwde antwoorden biedt.

Dit onderzoek richt zich op het verbeteren van het Context-Aware Knowledge Extraction (CAKE)-framework, een RAG-systeem dat videobeelden verwerkt. De huidige visuele informatie-extractie is beperkt, daarom wordt een nieuwe methode ontwikkeld waarin objectdetectie en Human Action Recognition (HAR) worden gecombineerd. De gedetecteerde objecten en acties worden gekoppeld op basis van tijd, ruimtelijke nabijheid en semantische relevantie.

De methode wordt geëvalueerd op modelprestaties, prestaties van het CAKE-framework en toepasbaarheid in industriële trainingen. De resultaten zullen mogelijk bijdragen aan een verbeterde verwerking van visuele informatie en de effectiviteit van virtuele les assistenten mogelijk verhogen, waardoor bedrijven hun werknemers beter kunnen ondersteunen.

Index Terms—Retrieval-Augmented Generation (RAG), Large Language Model (LLM), Context-Aware Knowledge Extraction (CAKE), Objectdetectie, Human Action Recognition (HAR), Knowledge Graphs, Virtuele les assistent

I. INTRODUCTIE

Uit onderzoek blijkt dat 42% van de bedrijven in de industrie te kampen heeft met een personeelstekort [40]. De verwachting is dat dit tekort binnen twee jaar zal oplopen tot 69% [40]. Daarnaast ervaart 22,4% van de bedrijven hinder bij het vinden van personeel vanwege een beperkt aantal werkzoekenden [2]. Voor 25,9% vormt een gebrek aan scholing, kwalificaties of relevante ervaring bij potentiële werknemers een belangrijke belemmering [2]. Om dit probleem te overvangen, investeert 24,4% van de bedrijven in trainingen en ontwikkelingsprogramma's voor hun medewerkers [4].

JOZ, een fabrikant van landbouwmachines, ervaart vergelijkbare uitdagingen op het gebied van personeelstekort en werving. Om de vaardigheden van werknemers te verbeteren, biedt JOZ eveneens online trainingen aan. Hoewel dit een stap in de juiste richting is, ontbreekt er interactiviteit aan deze trainingen. Werknemers hebben geen mogelijkheid om vragen te stellen of ondersteuning te krijgen, wat hun leerproces belemmert.

Een effectieve leeroplossing zou niet alleen toegang tot relevante informatie moeten bieden, maar ook interactieve ondersteuning, zoals het beantwoorden van vragen op basis van de aangeboden lesstof. Traditionele chatbots zijn hiervoor vaak beperkt, omdat ze generieke antwoorden geven en geen specifieke kennis over de training hebben.

Een mogelijke oplossing is het gebruik van een Retrieval-Augmented Generation (RAG)-systeem. Dit systeem maakt gebruik van een Large Language Model (LLM) dat relevante informatie uit de lesstof ophaalt en verwerkt in zijn antwoorden. Hierdoor kunnen werknemers op een natuurlijke manier vragen stellen en direct ondersteuning krijgen bij hun leerproces. Onderzoek toont aan dat RAG-systemen effectief kunnen zijn bij het verbeteren van gepersonaliseerd leren en contextspecifieke ondersteuning [6], [17], [30].

RAG haalt informatie gebruikelijk uit documenten. Veel online trainingen en lessen worden echter aangeboden in de vorm van video's. Hierdoor is een traditioneel RAG-systeem, dat zijn antwoorden baseert op documenten, niet voldoende.

Om de beperkingen van traditionele RAG-systemen te verhelpen, is een oplossing ontwikkeld: het Context-Aware Knowledge Extraction (CAKE) framework. Dit framework is ontworpen om contextuele kennis te extraheren en te benutten uit diverse bronnen, zoals video's, audio en tekst. Door gebruik te maken van Knowledge Extraction-algoritmes, verbetert CAKE de prestaties van Large Language Models (LLM's) door relevante informatie uit multimodale gegevens te halen en effectief in te zetten [43] (NOG EEN BRON TOEVOEGEN).

Uit het onderzoek blijkt echter dat de huidige methode binnen het CAKE-framework voor het extraheren van informatie uit videobeelden niet optimaal werkt. Daarom wordt in dit onderzoek een nieuwe methode voorgesteld, waarbij gebruik wordt gemaakt van een objectdetectie- en een Human Action Recognition (HAR) model. Vervolgens wordt een methode ontwikkeld die de output van beide modellen combineert. Deze gegevens worden vervolgens opgeslagen in een knowledge graph en geïntegreerd in het bestaande CAKE-framework.

Het doel van dit onderzoek is om de effectiviteit van RAG-systemen bij het verwerken van videobeelden te verbeteren. Door objectdetectie en HAR te combineren en de verkregen informatie op te slaan in een knowledge graph binnen het CAKE-framework, kunnen virtuele les assistenten binnen online trainingen en lessen in de industrie effectiever ondersteunen.

Om dit doel te bereiken, wordt de volgende onderzoeks-vraag onderzocht: *Hoe kunnen objectdetectie en Human Action Recognition worden ingezet om informatie uit videobeelden te extraheren en te integreren in een knowledge graph binnen het CAKE-framework om de effectiviteit van virtuele les assistenten in online trainingen en lessen in de industrie te verbeteren?* Deze onderzoeks-vraag wordt opgedeeld in de volgende deelvragen:

- Hoe kunnen alleen objecten worden opgeslagen in de knowledge graph die relevant zijn voor de online trainingen en lessen?*
- Hoe kan het gedetecteerde object effectief gekoppeld worden aan de uitgevoerde actie?*
- Hoe kan de rekenintensiteit van de objectdetectie en Human Action Recognition worden beperkt?*
- Wat is de toegevoegde waarde van de virtuele les assistent vergeleken met de huidige situatie?*

In hoofdstuk 2 wordt het theoretisch kader opgesteld. In hoofdstuk 3 wordt de voorgestelde onderzoeks-methode besproken. In hoofdstuk 4 staat de planning van het onderzoek centraal. Tot slot gaat hoofdstuk 5 in op de conclusie.

II. REQUIREMENTS

JOZ heeft de volgende requirements opgegeven waar het volledige RAG-framework aan moet voldoen.

III. THEORETISCHE KADER

Dit onderzoek maakt deel uit van een groter project waarin informatie uit video's wordt opgeslagen in een knowledge graph en vervolgens wordt gebruikt in een Retrieval-Augmented Generation (RAG) systeem. Dit systeem maakt gebruik van een Large Language Model (LLM) om antwoorden te genereren op basis van gestructureerde kennis.

Om de basis van dit grotere onderzoek te begrijpen, worden in dit theoretisch kader de relevante concepten besproken. Eerst worden Large Language Models (LLM's), Retrieval-Augmented Generation (RAG) en knowledge graphs uitgelegd. Vervolgens wordt ingegaan op Graph Retrieval-Augmented Generation (GRAG), een techniek die knowledge graphs integreert in RAG-systemen. Daarna wordt het CAKE-framework beschreven waarin video's worden gebruikt in een GRAG-pipeline. Tot slot worden objectdetectie, Human Action Recognition (HAR) en Human-Object Interaction (HOI) besproken, drie technieken die essentieel zijn voor het extraheren van betekenisvolle informatie uit videobeelden.

A. Large Language Models (LLM's)

Large Language Models (LLM's) zijn modellen die zijn getraind op het genereren van menselijke taal [48]. Door het trainen op grote hoeveelheden tekstdaten leren deze modellen complexe taalpatronen te herkennen en kunnen ze diverse taalgerelateerde taken uitvoeren [31].

De meeste LLM's zijn gebaseerd op de transformer-architectuur [11]. Deze architectuur maakt gebruik van het self-attention mechanism, waardoor modellen effectief verbanden tussen ver uit elkaar liggende woorden in de tekst kunnen

Requirement	Prioriteit	Omschrijving	Onderbouwing
Het model beantwoordt hele technische vragen.	Must	De video's van JOZ bevatten domeinspecifieke technische informatie. Het model moet in staat zijn om vragen hierover correct te beantwoorden.	JOZ-video's bevatten technische informatie, en het model moet deze correct kunnen verwerken om gebruikers goed te ondersteunen. Dit voorkomt misinterpretaties, verhoogt de betrouwbaarheid en zorgt voor efficiënte kennisoverdracht binnen het domein.
Het model moet Engels, Nederlands, Duits, Frans en Italiaans spreken.	Must	JOZ heeft meerdere vestigingen in het buitenland. Daarom is het belangrijk dat het model minimaal in deze talen kan antwoorden.	JOZ heeft vestigingen in verschillende landen. Het ondersteunen van deze talen is belangrijk om effectief te communiceren met monteurs en misverstanden te voorkomen.
Makkelijk te trainen of hertrainen.	Must	Wanneer er een nieuwe video wordt opgenomen over een nieuw apparaat, moet het model eenvoudig opnieuw getraind kunnen worden.	JOZ blijft haar producten uitbreiden. Nieuwe apparaten betrekken nieuwe trainingsvideo's. Het model moet snel en zonder zware computers hertraind kunnen worden.
Hoge betrouwbaarheid.	Must	Fouten kunnen negatieve gevolgen hebben voor de opleiding van monteurs en kunnen leiden tot fouten bij klanten.	Onbetrouwbare informatie kan leiden tot defecten bij klanten, financiële schade, veiligheidsrisico's en verlies van klantvertrouwen.
Antwoorden moeten gerelateerd zijn.	Must	De antwoorden van het model moeten altijd gerelateerd zijn aan de inhoud van de video's.	Irrelevante antwoorden kunnen verwarring veroorzaken en fouten in de praktijk opleveren. Dit vermindert de efficiëntie van de training.
De antwoorden moeten binnen 1-3 seconden starten met genereren.	Should	Het genereren van antwoorden moet snel beginnen.	Lange wachttijden ontmoedigen gebruikers om het systeem te gebruiken.
Het model moet Tsjechisch, Russisch, Pools, Spaans, Japans, Zuid-Koreans en Noors spreken.	Should	JOZ heeft veel vestigingen in het buitenland. Het is daarom gewenst dat het model ook deze talen ondersteunt.	Ondersteuning van extra talen helpt lokale technici en monteurs, waardoor communicatiebarrières worden verlaagd.
Vragen en antwoorden inzien en daarmee het model verbeteren.	Could	Bij analyse moet het mogelijk zijn om het model gericht te verbeteren.	Het kunnen inzien en analyseren van vragen en antwoorden biedt waardevolle inzichten om het model gericht te verbeteren, bijvoorbeeld door het identificeren van fouten, onnauwkeurigheden of terugkerende patronen. Dit helpt bij optimalisatie en verhoogt de betrouwbaarheid van het model.
Antwoorden opzoeken op het internet.	Won't	Vanwege de specifieke domeinkennis zal het model geen gebruik maken van informatie uit externe bronnen op het internet.	Externe bronnen kunnen leiden tot foutieve antwoorden die niet aansluiten bij de verwachte JOZ-standaarden.

Tabel I
OVERZICHT VAN REQUIREMENTS

vastleggen [45]. Sinds de introductie van de transformer-architectuur zijn verschillende LLM's ontwikkeld, zoals GPT (Generative Pre-trained Transformer) en BERT (Bidirectional Encoder Representations from Transformers).

LLM's hebben potentie om ingezet te worden in het onderwijs. Ze kunnen ondersteuning bieden bij het ontwikkelen van lees-, schrijf-, wiskunde- en taalvaardigheden. Daarnaast kunnen ze studenten voorzien van gepersonaliseerd oefenmateriaal, samenvattingen en uitleg, wat kan bijdragen aan betere leerprestaties en een verrijkte leerervaring [22].

B. Retrieval Augmented Generation (RAG)

LLM's worden getraind op grote, algemene datasets en beschikken daardoor over weinig tot geen domeinspecifieke kennis [5]. Om LLM's toch toegang te geven tot domeinspecifieke kennis, kan Retrieval Augmented Generation (RAG) worden ingezet. Daarnaast zorgt RAG ervoor dat LLM's minder hallucineren [25], wat een voordeel met zich meebrengt binnen het onderwijs. RAG kan worden onderverdeeld in de volgende basiscategorieën [14]:

1) *Naive RAG*: Naive RAG is de eenvoudigste vorm van RAG. Hierbij wordt tekst opgedeeld in kleinere stukken tekst, deze kleinere stukken tekst worden chunks genoemd. Deze chunks worden met een embedding model omgezet in vectorrepresentaties en vervolgens opgeslagen in een vectordatabase. Wanneer een gebruiker een vraag stelt aan het LLM, wordt deze vraag ook omgezet in een vectorrepresentatie. Vervolgens wordt deze vector vergeleken met alle vectoren in de vectordatabase. De K meest overeenkomende chunks worden opgehaald uit de vectordatabase en toegevoegd aan de prompt als context. Op basis van deze extra context kan het LLM een beter onderbouwd antwoord genereren [14].

2) *Advanced RAG*: Advanced RAG is een uitbreiding ten opzichte van Naive RAG. Hierbij worden twee stappen toegevoegd. Er wordt een pre-retrieval stap toegevoegd om de kwaliteit van de geïndexeerde tekst te verbeteren en de oorspronkelijke vraag van de gebruiker duidelijker en beter geschikt te maken voor de retrieval-taak. Daarnaast wordt er een post-retrieval stap toegevoegd om een overload aan irrelevante informatie te voorkomen. Dit proces richt zich op het selecteren van essentiële informatie, het benadrukken van de belangrijke onderdelen en het verkorten van de context die verwerkt moet worden [14].

Naast deze twee basiscategorieën zijn er nog vele uitbreidings mogelijk op een RAG-framework.

Onderzoek toont aan dat het gebruik van RAG in het onderwijs diverse voordelen heeft. Zo blijkt uit het onderzoek van *Hicke et al.* [17] dat de inzet van RAG leidt tot een verbetering van 30% in de antwoordkwaliteit. Daarnaast toont het onderzoek van Alawwad et al. [6] aan dat de nauwkeurigheid van antwoorden op de testset met bijna 10% toeneemt door het gebruik van RAG. Verder geeft het onderzoek van *Ma et al.* [30] aan dat 78% van de gebruikers aangaven dat AI lesassistenten hun leerproces verbeterden.

Samengevat laat onderzoek zien dat RAG inzetten in het onderwijs voordelen biedt. LLM's kunnen beter domeinspecifieke antwoorden geven en minder hallucinaties vertonen. Studies laten zien dat de kwaliteit van antwoorden omhooggaat en dat gebruikers ervaren dat hun leerproces verbetert.

C. Knowledge graphs

In plaats van vectordatabases kan RAG ook gebruikmaken van knowledge graphs. Knowledge graphs zijn structuren die informatie organiseren over entiteiten zoals objecten, gebeurtenissen, situaties en concepten. Ze worden gebruikt om complexe netwerken van gerelateerde informatie te modelleren, waardoor ze gebruikt kunnen worden om kennis uit af te leiden en over te dragen [18].

Er is geen universeel geaccepteerde definitie van knowledge graphs. In het paper '*Knowledge graphs*' wordt de volgende inclusieve definitie gehanteerd: *Knowledge graphs worden beschouwd als een grafiek van gegevens die bedoeld is om kennis over de echte wereld te verzamelen en over te dragen, waarbij de knooppunten belangrijke entiteiten vertegenwoordigen en de randen de relaties tussen deze entiteiten weergeven* [18].

Het ontwikkelen van knowledge graphs wordt doorgaans onderverdeeld in de volgende fases:

- 1) Specificatiefase: In deze fase wordt de algemene motivatie, doel en het onderwerp voor het maken van de knowledge graph vastgesteld.
- 2) Conceptualisatiefase: In deze fase wordt de verzamelde domein kennis gestructureerd in de vorm van tussen representaties.
- 3) Formalisatiefase: In deze fase wordt het conceptuele model van het domein omgezet in een, door machines interpreteerbaar model.
- 4) Integratiefase: In deze fase wordt het model uitgebreid met bestaande kennis.
- 5) Augmentatiefase: In deze fase wordt de kwaliteit beoordeelt en de knowledge base verfijnd.

Met deze algemene stappen kan een knowledge graph worden opgebouwd [20]. Deze informatie is belangrijk voor het begrijpen van een knowledge graph binnen het CAKE-framework.

D. Graph Retrieval-Augmented Generation (GRAG)

Graph Retrieval-Augmented Generation (GRAG) is een variant op RAG waarbij gebruikgemaakt wordt van knowledge graphs. RAG werkt zoals eerder uitgelegd door documenten op te halen op basis van tekstuele overeenkomst, maar houdt geen rekening met de onderlinge relaties van de chunks. GRAG houdt hier wel rekening mee door gebruik te maken van knowledge graphs [19].

Tijdens de retrieval fase identificeert GRAG eerst de meest relevante knooppunten in de knowledge graph en selecteert hun directe buren. Vervolgens wordt met een soft pruning techniek irrelevante informatie verwijderd uit de subgraph [19].

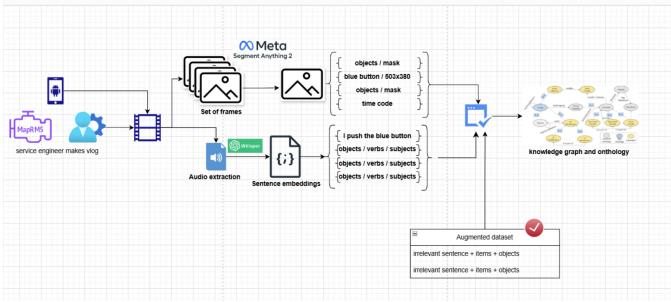
De subgraph wordt vervolgens op twee manieren verwerkt door het LLM. De eerste manier is de graph view, waarbij representaties van de subgraphs worden geleerd als soft prompt om vast te leggen hoe teksten met elkaar verbonden zijn. De tweede manier is text view, waarbij de subgraphs worden omgezet in hiërarchische tekstbeschrijvingen als hard prompt om vast te leggen hoe de verbindingen worden verwoord [19].

Uit de resultaten blijkt dat GRAG beter presteert dan andere RAG-methoden en zelfs gefinetunedde LLM's overtreft op alle metrics en datasets [19]. Dit laat dus zien dat GRAG een effectievere methode is voor het verbeteren van LLM-prestaties.

E. Context-Aware Knowledge Extraction (CAKE)

In dit onderzoek wordt gewerkt aan de verbetering van het CAKE-framework. Dit framework is een GRAG-pipeline waarin een LLM antwoorden genereert op basis van video's. In Figuur 1 is het gedeelte van de pipeline weergegeven waarin de video's worden omgezet in een knowledge graph.

In dit proces worden de video's opgedeeld in audio en beeld. De audio wordt vervolgens getranscribeerd naar tekst met behulp van het wisper large-v2 model van openAI. De output is een tekstbestand. Deze tekst wordt vervolgens met



Figuur 1. Schematische weergave CAKE-framework.

SDPMChunker van Chonkie opgedeeld in kleinere chunks. De start- en eindtijd van wanneer deze chunk voorkomt in de video wordt bijgehouden. Vervolgens wordt deze informatie opgeslagen in een JSON-bestand. Nadat het audiodeel is uitgevoerd, wordt het beeld verwerkt. Het beeld van de video wordt op basis van de chunks die zijn gemaakt in het audiodeel gesegmenteerd. Vervolgens wordt in elk segment per drie seconden objecten gedetecteerd in het beeld met behulp van het YOLOv8-model. Deze informatie wordt opgeslagen in een JSON-bestand en wordt gekoppeld aan de bijbehorende chunk. Deze data wordt vervolgens omgezet in een knowledge graph [43], [44] (NOG 1 BRON TOEVOEGN).

Uit de resultaten blijkt dat het extraheren van de juiste informatie niet altijd voldoende werkt. Er worden vaak te veel niet-relevante objecten gedetecteerd en daarnaast worden de objecten ook regelmatig onjuist gelabeld [44].

Aangezien in dit onderzoek gefocust wordt op het verbeteren van het extraheren van visuele informatie uit de video's, wordt er in het theoretisch kader verder ingegaan op objectdetectie en HAR.

F. Objectdetectie

Objectdetectie is een onderdeel van computervision en houdt zich bezig met het detecteren van objecten en een bepaalde klasse in afbeeldingen. Het doel van objectdetectie is het ontwikkelen van een model dat aangeeft welke objecten zich waar bevinden [49].

De meeste objectdetectiemodellen kunnen tegenwoordig ingedeeld worden in de klasse Two-stage detector en One-stage detector [49] [23].

De Two-stage detector verdeelt het detecteren van objecten in twee taken, het lokaliseren en het classificeren van het object. Eerst wordt de regio bepaald waar het object zich in bevindt en vervolgens wordt de regio geclasseerd. Het voordeel van deze categorie is de hoge nauwkeurigheid, het nadeel is de lage detectiesnelheid [23]. Voorbeelden van Two-stage detector modellen zijn: Faster RCNN [15] en FPB [29].

De One-stage detector lokaliseert en classificeert objecten tegelijkertijd, zonder ze in twee afzonderlijke stappen te verdelen [23]. Voorbeelden van One-stage detector modellen zijn: YOLO [36] en DETR [8].

Het YOLO framework is een veelgebruikt framework en staat bekend om zijn goede balans tussen snelheid en accuracy

[39]. Dit framework wordt ook vaak toegepast in de industrie [21], [35], [42].

Het recentste model in de YOLO-serie is YOLOv12. Voordien maakten YOLO-modellen gebruik van convolutionele neurale netwerken (CNN), omdat met attention-based modellen de snelheid van CNN-gebaseerde modellen niet kon worden geëvenaard. YOLOv12 brengt hier verandering in door wel gebruik te maken van een attention-based model, terwijl het dezelfde snelheid behaalt als CNN-gebaseerde YOLO-modellen. Bovendien levert YOLOv12 betere prestaties dan alle voorgaande YOLO-modellen [41].

Een bijkomend voordeel is dat YOLOv12 een relatief compact model is. Afhankelijk van de gekozen variant varieert het aantal parameters tussen 2,6 miljoen en 59,1 miljoen [41].

Kort samengevat toont onderzoek aan dat YOLOv12 een van de snelste en best presterende objectdetectiemodellen is, met als bijkomend voordeel dat het relatief compact is.

G. Human Action Recognition (HAR)

Human Action Recognition richt zich op het begrijpen en labelen van menselijke acties [38]. Dit proces herkent acties op basis van video's waarin de volledige handeling is vastgelegd. Dit verschilt van action prediction, waarbij de actie wordt voorspeld op basis van een onvolledige video [26].

De meest gebruikte data voor HAR is RGB-video. Dit datatype heeft als voordeel dat het breed toepasbaar is en eenvoudig te verkrijgen en te gebruiken, bijvoorbeeld bij lesvideo's. Echter, RGB-video heeft ook enkele nadelen. HAR-modellen die op deze data zijn getraind, zijn gevoelig voor factoren zoals kijkhoek, achtergrond en verlichting [38].

Het SlowFast Network is een veelgebruikt model voor HAR en is eerder ingezet in de industrie om menselijke acties te detecteren [27]. Zoals de naam al zegt, maakt dit model gebruik van een langzaam pad en een snel pad. Het langzame pad is ontworpen om informatie vast te leggen uit enkele frames. Dit pad gebruikt een lage framerate en een trage vernieuwingssnelheid. Het snelle pad is ontworpen om snel veranderende bewegingen vast te leggen door te werken met een hoge vernieuwingssnelheid [13].

Een ander veelgebruikt model is Temporal Shift Module (TSM) [28]. Traditionele 2D CNN's verwerken afzonderlijke frames en kunnen daardoor geen tijdsgebonden relaties herkennen. TSM lost dit probleem op door een deel van de kanalen in de tijdsdimensie naar voren en naar achteren te verschuiven. Hierdoor wordt informatie uit naburige frames geïntegreerd in het huidige frame. Dit model behaalt de prestaties van een 3D CNN, terwijl het de complexiteit van een 2D CNN behoudt [28].

H. Human-Object Interaction (HOI)

Human-Object Interaction (HOI)-detectie is een combinatie van objectdetectie en HAR. Het doel van HOI-detectie is om mensen en objecten in afbeeldingen of video's te lokaliseren en vervolgens de interacties tussen hen te detecteren. HOI-detectiemodellen zijn grofweg te verdelen in twee categorieën: two-stage en one-stage detectie [46].

Two-stage modellen genereren eerst voorstellen voor regio's waarin mensen en objecten voorkomen. Vervolgens worden deze voorstellen verwerkt door een relationship-classificatiennetwerk, dat op basis van kenmerken van mensen en objecten interacties classificeert [46]. Een voorbeeld van een two-stage model is ViPLO, dat gebruikmaakt van een Vision Transformer (ViT) backbone in combinatie met een Graph Neural Network [34]. ViPLO behaalt state of the art prestaties [34] en is getraind op vier RTX 3090 GPUs met 24 GB [33].

One-stage modellen voorspellen de interacties tussen mensen en objecten direct in één keer [34]. Een voorbeeld hiervan is HOTR [24], een op transformers gebaseerd model dat interacties in één keer classificeert. HOTR behaalt state-of-the-art prestaties [24].

IV. ONDERZOEKSMETHODE

Uit onderzoek blijkt dat het extraheren van informatie uit het beeld van de video's in het CAKE-framework nog onvoldoende werkt, waardoor de knowledge graph onvolledig blijft [44]. Dit onderzoek richt zich op het verbeteren van het extraheren van informatie uit de beelden van de video's.

Er is in dit onderzoek voor gekozen om geen gebruik te maken van Human-Object Interaction (HOI), maar in plaats daarvan objectdetectie te combineren met Human Action Recognition (HAR). De keuze is gebaseerd op de volgende redenen:

Ten eerste blijkt uit het literatuuronderzoek dat HOI-modellen erg rekenintensief zijn. Het ViPLO-model vereist bijvoorbeeld vier RTX 3090 GPU's met 24 GB aan geheugen. Een belangrijke requirement van dit onderzoek is dat het model eenvoudig getraind kan worden, bijvoorbeeld wanneer een nieuwe video met een nieuw apparaat wordt toegevoegd. Met zulke zware modellen is dit niet haalbaar.

Ten tweede zijn HOI-modellen gericht op het detecteren van interacties tussen mensen en objecten. Dit kan echter een beperking vormen, omdat sommige objecten ook van belang kunnen zijn zonder dat er een interactie met een persoon plaatsvindt. Door objectdetectie te combineren met HAR, kan waardevolle informatie uit videobeelden worden gehaald, ongeacht of er interactie tussen mens en object is.

Deze aanpak biedt meer flexibiliteit en schaalbaarheid, terwijl de requirement van rekenintensiviteit beheersbaar blijft.

A. Objectdetectie

Om het huidige proces van het extraheren van informatie uit het beeld van video's te verbeteren, wordt de volgende oplossing voorgesteld. In de huidige methode wordt elke drie seconden op één frame objectdetectie uitgevoerd met het YOLOv8-model. Zoals al eerder genoemd, is uit eerder onderzoek echter gebleken dat deze aanpak niet optimaal presteert.

Om de prestaties van het objectdetectiemodel te verbeteren, wordt voorgesteld om een pretrained YOLOv12-model [41] te gebruiken en dit specifiek te fine-tunen op de meest voorkomende objecten in de video's van JOZ. Hierbij is het

belangrijk dat het model zijn bestaande kennis van andere objecten behoudt, zodat het niet alleen de nieuwe objecten leert herkennen, maar ook zijn oorspronkelijke detectiemogelijkheden behoudt. Dit is nuttig, omdat de bestaande kennis van YOLOv12 mogelijk relevant kan zijn voor het detecteren van andere algemene objecten in de video's.

Tijdens het fine-tunen worden extra klassen toegevoegd die het model moet leren herkennen. Deze nieuwe klassen zijn afkomstig uit de data van JOZ.

Bij het fine-tuneproces worden op zijn minst de volgende hyperparameters geoptimaliseerd om het best presterende model te verkrijgen:

- Epochs: Deze hyperparameter wordt aangepast omdat meer epochs het model helpen patronen beter te leren, maar te veel epochs kunnen leiden tot overfitting.
- Learning rate: Deze hyperparameter wordt aangepast omdat een te hoge waarde kan leiden tot instabiele training en een te lage waarde tot langzaam leren.
- Batch size: Deze hyperparameter wordt aangepast omdat een grotere batchgrootte de training versnelt, maar mogelijk meer rekenkracht vereist en de generaliseerbaarheid kan beïnvloeden.

Er is voor gekozen om YOLOv12 te gebruiken, omdat dit een van de snelste en best presterende objectdetectiemodellen is. Daarnaast is het relatief klein, wat het makkelijker maakt als het model opnieuw getraind moet worden.

Na de optimalisatie van het objectdetectiemodel wordt de aanpak verbeterd door gebruik te maken van Multi-Frame Tracking, in plaats van objectdetectie uit te voeren op één frame per drie seconden. Multi-Frame Tracking zorgt ervoor dat objecten in opeenvolgende frames niet telkens als nieuwe entiteiten worden behandeld. Zonder deze techniek zou een object in elk frame opnieuw worden gedetecteerd zonder een uniek ID. Door Multi-Frame Tracking behoudt een herkend object gedurende meerdere frames hetzelfde ID, wat de nauwkeurigheid verhoogt en ruis verminderd. Objectdetectie kan namelijk soms objecten missen in een frame, maar Multi-Frame Tracking kan deze gaten opvullen en zorgt voor een consistent resultaat (BRON NOG TOEVOEGEN).

Daarnaast zorgt Multi-Frame Tracking voor een efficiëntere verwerking. Doordat objecten al een ID hebben, hoeft de detectie alleen te controleren of ze al bestaan, in plaats van ze telkens opnieuw te analyseren. Dit vermindert de rekenkosten en versnelt het proces aanzienlijk (BRON NOG TOEVOEGEN).

Een ander voordeel van Multi-Frame Tracking is dat het herhaling bij het opslaan voorkomt. Aangezien acties meestal langer duren dan één frame, zou het inefficiënt zijn om per frame een object-actie relatie op te slaan. Multi-Frame Tracking zou hierbij kunnen bijdragen door bijvoorbeeld bij te houden wanneer een object-actie relatie begint en eindigt. Hierdoor wordt er niet onnodig dubbele data opgeslagen in de knowledge graph.

B. Human Action Recognition (HAR)

In het huidige proces wordt er alleen rekening gehouden met de objecten die in het beeld staan. Echter, is het ook belangrijk welke handelingen er met deze objecten worden uitgevoerd. Daarom wordt er voorgesteld om naast objectdetectie gebruik te maken van Human Action Recognition (HAR).

Voor HAR wordt het TSM-model gebruikt, dat is getraind op de Something-Something v2-dataset. Dit model is gekozen vanwege de goede prestaties en lage rekentijd.

De Something-Something v2-dataset, waar het model op getraind is, bevat video's van fysieke handelingen en interacties met objecten [16]. In totaal omvat de dataset 220.847 video's met 174 unieke labels [1]. Enkele voorbeelden van deze labels zijn:

- “Putting something next to something”
- “Covering something with something”
- “Pushing something from left to right”

Door de specifieke structuur van deze labels is een model dat op deze dataset is getraind goed te combineren met een objectdetectiemodel. De *something* kan als placeholder dienen voor het daadwerkelijke object.

C. Combinatie objectdetectie en HAR

Om de gedetecteerde objecten en menselijke acties efficiënt op te slaan in de knowledge graph zal er een koppeling worden gemaakt tussen het gedetecteerde object en de menselijke actie. Hierbij moet rekening worden gehouden dat de juiste actie aan het juiste object wordt gekoppeld. Ook zullen er objecten zijn die wel relevant zijn, maar waarmee geen actie wordt uitgevoerd.

Om dit uit te voeren zal een post-hoc koppeling worden toegepast, waarbij het objectdetectie- en HAR-model eerst afzonderlijk een voorspelling genereren. Deze voorspellingen worden vervolgens achteraf gekoppeld, waarbij rekening wordt gehouden met het tijdscomponent, de ruimtelijke nabijheid en de semantische relevantie.

Het eerste onderdeel waarmee rekening moet worden gehouden, is tijd. Om objecten correct aan een actie te koppelen, moeten ze gedeeltelijk overlappen in de tijd. Bijvoorbeeld, als een object zichtbaar is van seconde 0 tot 5 en een actie wordt gedetecteerd van seconde 4 tot 10, dan is er een gedeeld tijdsinterval tussen seconde 4 en 5. Om deze overlap te bepalen, wordt zoals eerder aangegeven, gebruikgemaakt van Multi-Frame Tracking. Het model dat hiervoor wordt ingezet, is DeepSORT [47]. DeepSORT is een uitbreiding van Simple Online and Realtime Tracking (SORT) [7] die gebruikmaakt van een CNN [47]. Met dit model wordt voor elk object een unieke ID toegekend en wordt bijgehouden welk object en op welke frames het zichtbaar was.

Als er een tijdsinterval is waarin zowel het object als de actie voorkomen, wordt binnen dit interval de ruimtelijke nabijheid onderzocht. Dit gebeurt door de gemiddelde Intersection over Union (IoU) te berekenen tussen de bounding boxes van het object en de actie gedurende het hele interval. Wanneer deze gemiddelde IoU een nog te bepalen drempel overschrijdt, is de

kans groot dat het object daadwerkelijk bij de actie betrokken is.

Tot slot, wanneer gemiddelde IoU boven de drempel uitkomt, wordt een semantische controle uitgevoerd om onlogische combinaties te voorkomen. Er wordt met Sentence-BERT [37] vectorrepresentaties gegenereerd van het gedetecteerde object en de actie. Vervolgens wordt de cosinusgelijkenis berekend tussen de embeddings van beide, om te bepalen in hoeverre ze semantisch verwant zijn. Wanneer deze gelijkenis een nog te bepalen threshold overschrijdt, wordt de koppeling tussen het object en de actie opgeslagen in de knowledge graph.

Voor objecten die wel worden gedetecteerd, maar waarbij geen actie wordt uitgevoerd, wordt gecontroleerd of ze in de audio worden genoemd. Binnen het CAKE-framework wordt de audio al getranscribeerd, waardoor kan worden gecontroleerd of het gedetecteerde object in de transcriptie voorkomt. Als het object in de transcriptie wordt vermeld, wordt het opgeslagen als een passief object. Wordt het object niet genoemd, dan wordt het als irrelevant beschouwd en niet opgeslagen.

D. Data

De input voor het volledige framework bestaat uit training- en instructievideo's van JOZ die worden gebruikt in hun online leeromgeving. Op het moment van schrijven is er echter nog geen toegang tot deze video's. Wel is toegang verleend tot video's van de Japanse vestigingen van het bedrijf, die vergelijkbaar zijn met de trainings- en instructievideo's.

In Figuur 2 en 3 zijn screenshots uit deze video's te zien, waarin instructies worden gegeven over de reparatie van de drivemotor en het mainboard. Het is belangrijk dat het uiteindelijke model in staat is om de relevante onderdelen en de bijbehorende acties te detecteren.



Figuur 2. Screenshot uit de video over de drivemotor.

Omdat het om zeer specifieke onderdelen gaat, zal een YOLOv12-model worden gefinetuned op deze machineonderdelen. Hiervoor wordt bij voorkeur een dataset van JOZ gebruikt. Mocht de kwaliteit van deze dataset onvoldoende zijn, dan wordt een eigen dataset samengesteld met foto's van de onderdelen die JOZ als belangrijk beschouwt.



Figuur 3. Screenshot uit de video over het mainboard.

E. Evaluatie

Het model wordt op drie manieren geëvalueerd. Ten eerste wordt het objectdetectiemodel afzonderlijk beoordeeld. Daarnaast wordt het volledige RAG-framework geëvalueerd op prestaties en effectiviteit in de praktijk.

1) *Evaluatie objectdetectie:* Voor de evaluatie van het gefinetunedede YOLOv12-model zal er gebruikgemaakt worden van Mean Average Precision (mAP). Deze metric is gekozen omdat mAP een van de meest gebruikte metrics binnen objectdetectie is [32]. De mAP is het gemiddelde van de Average Precision (AP) over alle klassen [32]. De AP is gebaseerd op het oppervlak onder de precision-recall curve [32].

2) *Evaluatie CAKE-framework:* Naast de evaluatie van het objectdetectiemodel zelf, wordt ook het volledige CAKE-framework geëvalueerd. Er wordt een baseline vastgesteld met het huidige framework, die vervolgens wordt vergeleken met het aangepaste framework waarin de informatie-extractiemethode uit dit onderzoek is geïmplementeerd. Dit maakt het mogelijk om te bepalen of de verbeteringen in het objectdetectiemodel daadwerkelijk toegevoegde waarde hebben voor het CAKE-framework.

Voor de evaluatie van het CAKE-framework zal Retrieval Augmented Generation Assessment (RAGAs) [10] worden gebruikt. RAGAs is een verzameling LLM-gebaseerde metrics voor het beoordelen van RAG-pipelines en vereist niet altijd het gebruik van een referentieantwoord [10].

De eerste metric die wordt gebruikt, is answer correctness. Deze metric bepaalt de correctheid van het antwoord door de ground truth te vergelijken met het door het LLM gegenereerde antwoord. Hiervoor worden de volgende categorieën geïdentificeerd:

- True Positive (TP): Statements die zowel in de ground truth als in het gegenereerde antwoord aanwezig zijn.
- False Positive (FP): Statements die in het gegenereerde antwoord voorkomen, maar niet in de ground truth.
- False Negative (FN): Statements die in de ground truth aanwezig zijn, maar ontbreken in het gegenereerde antwoord.

Op basis hiervan wordt de F1-score berekend, zoals weergegeven in Formule 1 [3].

$$F_1 = \frac{|TP|}{|TP| + 0.5 \cdot (|FP| + |FN|)} \quad (1)$$

Deze metric is gekozen omdat hiermee de correctheid van de antwoorden automatisch kan worden beoordeeld.

Een andere metric uit RAGAs die wordt gebruikt, is faithfulness. Deze metric beoordeelt de mate van hallucinaties in het model door de opgehaalde context te vergelijken met het door het LLM gegenereerde antwoord. Faithfulness wordt berekend met Formule 2 [10].

$$F = \frac{\text{Aantal claims gesupport door de context}}{\text{Totaal aantal claims}} \quad (2)$$

Deze metric is geselecteerd om te bepalen of het model zich daadwerkelijk baseert op de video's, of dat het informatie hallucineert.

De laatste metric uit RAGAs die wordt gebruikt, is answer relevance. Deze metric meet hoe relevant het gegenereerde antwoord is voor de gestelde vraag. Answer relevance wordt berekend door nieuwe vragen te genereren op basis van het gegeven antwoord. Vervolgens wordt de cosinusgelijkenis berekend tussen deze gegenereerde vragen en de originele vraag. Tot slot wordt het gemiddelde van deze cosinusgelijkenissen bepaald [10].

Deze metric is gekozen om te achterhalen of de antwoorden relevanter zijn geworden in vergelijking met de baseline.

Uit onderzoek blijkt dat beoordeling door mensen de meest betrouwbare evaluatiemethode is [12]. Daarom wordt, naast de automatische metrics, ook human evaluation uitgevoerd. Hierbij worden een x aantal vragen opgesteld met betrekking tot de video's, die vervolgens worden beantwoord met behulp van het CAKE-framework. Een expert/monteur met domeinkennis over de specifieke video's beoordeelt de antwoorden aan de hand van de beoordelingscriteria in tabel 2, die is gebaseerd op het artikel *A Survey on Evaluation of Large Language Models* [9].

Beoordelingscriteria	Betekenis	Schaal	Vraag aan beoordelaar
Nauwkeurigheid	Beoordeelt hoe correct een antwoord is.	1 (zeer onnauwkeurig) - 5 (zeer nauwkeurig)	Hoe feitelijk correct is dit antwoord?
Relevantie	Meet hoe goed het gegenereerde antwoord aansluit bij de vraag.	1 (niet relevant) - 5 (zeer relevant)	Hoe goed sluit dit antwoord aan bij de vraag?
Vloeiendheid	Evalueert de grammaticale correctheid en de leesbaarheid van de tekst.	1 (niet vloeiend) - 5 (zeer vloeiend)	Hoe vloeiend en begrijpelijk is dit antwoord geschreven?
Transparantie	Beoordeelt hoe duidelijk het model zijn beslissingen en redeneringen communiceert.	1 (niet transparant) - 5 (zeer transparant)	Hoe duidelijk communiceert het model het redeneerproces en beslissingen?
Veiligheid	Controleert of de gegenereerde inhoud geen schadelijke of ongewenste informatie bevat.	1 (onveilig) - 5 (zeer veilig)	Bevat dit antwoord potentieel schadelijke of ongewenste informatie?
Human Alignment	Meet in hoeverre de inhoud overeenkomt met waarden en sociale normen.	1 (niet afgestemd) - 5 (zeer goed afgestemd)	Hoe goed komt het antwoord overeen met menselijke waarden en normen?

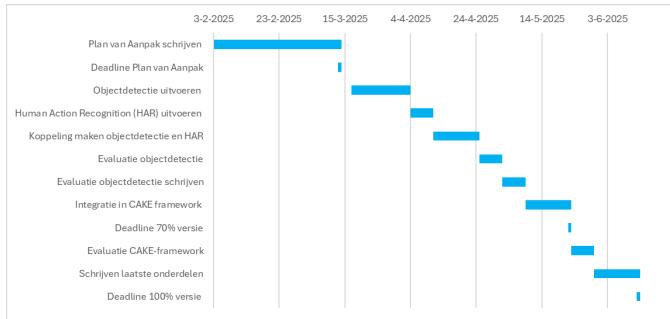
Tabel II
BEOORDELINGSCRITERIA VOOR GEGENEREERDE ANTWOORDEN.

3) *Evaluatie met gebruikers:* Naast de prestaties van het objectdetectiemodel en het CAKE-framework wordt ook de effectiviteit van het CAKE-framework als ondersteuning voor online trainingen en lessen in de industrie geëvalueerd. Hiervoor worden interviews gehouden met monteurs die de online lessen volgen. Vervolgens wordt er aan hen gevraagd of zij een les kunnen uitvoeren waarbij zij de mogelijkheid hebben tot ondersteuning van het CAKE-framework. Na afloop vindt

een tweede interview plaats om inzicht te krijgen in de toegevoegde waarde van het CAKE-framework en mogelijke verbeterpunten.

V. PLANNING

In Figuur 4 is de planning van de onderzoeksactiviteiten te zien.



Figuur 4. Planning onderzoeksactiviteiten.

VI. CONCLUSIE

Dit onderzoek richt zich op het verbeteren van de informatie-extractie uit videobeelden binnen het CAKE-framework, dat kan worden ingezet als een video-gebaseerde RAG-oplossing voor online trainingen en lessen in de industrie. Door objectdetectie en Human Action Recognition (HAR) te combineren, wordt een oplossing ontwikkeld die mogelijk de effectiviteit van virtuele les assistenten verhoogt.

De voorgestelde methode maakt gebruik van een YOLOv12-model, dat wordt gefinetuned op specifieke onderdelen en objecten uit de trainingsvideo's van JOZ. Daarnaast wordt een HAR-model geïmplementeerd om de uitgevoerde handelingen te herkennen. Vervolgens worden deze objecten en handelingen met elkaar gekoppeld, waarbij rekening wordt gehouden met tijd, ruimtelijke nabijheid en semantische relevantie. Hierdoor gaat mogelijk de betrouwbaarheid en nauwkeurigheid van de informatie-extractie omhoog.

De evaluatie van het model vindt plaats op drie niveaus: de prestaties van het objectdetectiemodel, de prestaties van het CAKE-framework als geheel en de praktische toepasbaarheid binnen de industriële trainingsomgeving. Hiervoor worden zowel automatische evaluatiemethoden, zoals RAGAs-metrics, als human evaluations ingezet. Daarnaast wordt de meerwaarde van de virtuele lesassistent beoordeeld door middel van interviews onder monteurs die de online trainingen volgen.

Deze aanpak zal mogelijk leiden tot een verbetering in de nauwkeurigheid en relevantie van de gegenereerde antwoorden, terwijl de rekenintensiteit beheersbaar blijft. Hierdoor kunnen werknemers in de industrie effectiever worden ondersteund in hun leerproces, wat bijdraagt aan een verbeterde kennisoverdracht.

REFERENTIES

- [1] Something-Something v. 2.
- [2] Hogere werkdruk belangrijkste gevolg personeeltekort volgens ondernemers, August 2024.
- [3] Answer correctness - Ragas, January 2025.
- [4] Centraal Bureau voor de . Drie kwart van de ondernemers probeert productiviteit te verhogen, May 2024.
- [5] Ankush Agarwal, Sakharam Gawade, Amar Prakash Azad, and Pushpak Bhattacharyya. KITLM: Domain-Specific Knowledge InTegration into Language Models for Question Answering, August 2023. arXiv:2308.03638 [cs].
- [6] Hessa A. Alawwad, Areej Allothali, Usman Naseem, Ali Alkhathlan, and Amani Jamal. Enhancing textual textbook question answering with large language models and retrieval augmented generation. *Pattern Recognition*, 162:111332, June 2025.
- [7] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, September 2016. arXiv:1602.00763 [cs].
- [8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers, May 2020. arXiv:2005.12872 [cs].
- [9] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A Survey on Evaluation of Large Language Models, December 2023. arXiv:2307.03109 [cs].
- [10] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *Association for Computational Linguistics*, pages 150–158, March 2024.
- [11] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–25, October 2024.
- [12] Amer Faraa, Zhen Yang, Kien Duong, Nadeesha Perera, and Frank Emmert-Streib. Evaluation of Question Answering Systems: Complexity of judging a natural language, September 2022. arXiv:2209.12617 [cs].
- [13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, Seoul, Korea (South), October 2019. IEEE.
- [14] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, March 2024. arXiv:2312.10997 [cs].
- [15] Ross Girshick. Fast R-CNN. pages 1440–1448. IEEE, 2015.
- [16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, June 2017. arXiv:1706.04261 [cs].
- [17] Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. AI-TA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs, December 2023. arXiv:2311.02775 [cs].
- [18] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge Graphs. *ACM Computing Surveys*, 54(4):1–37, May 2022. arXiv:2003.02320 [cs].
- [19] Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. GRAG: Graph Retrieval-Augmented Generation, October 2024. arXiv:2405.16506 [cs].
- [20] Ali Hur, Naeem Janjua, and Mohiuddin Ahmed. A Survey on State-of-the-art Techniques for Knowledge Graphs Construction and Challenges ahead. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 99–103, Laguna Hills, CA, USA, December 2021. IEEE.
- [21] Muhammad Hussain. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. *Machines*, 11(7):677, June 2023.

- [22] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tillman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, April 2023.
- [23] Ravpreet Kaur and Sarjeet Singh. A comprehensive review of object detection with deep learning. *Digital Signal Processing*, 132:103812, January 2023.
- [24] Bumssoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-End Human-Object Interaction Detection with Transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 74–83, Nashville, TN, USA, June 2021. IEEE.
- [25] Jason Kirchenbauer and Caleb Barns. Hallucination Reduction in Large Language Models with Retrieval-Augmented Generation Using Wikipedia Knowledge, May 2024.
- [26] Yu Kong and Yun Fu. Human Action Recognition and Prediction: A Survey, February 2022. arXiv:1806.11230 [cs].
- [27] Zhen Liang, Ying Hu, and Ransheng Yang. Research on Industrial Human Action Recognition based on Improved SlowFast. In *International Conference on Intelligent Autonomous Systems (ICOIAS)*, pages 94–99, 2023.
- [28] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal Shift Module for Efficient Video Understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, Seoul, Korea (South), October 2019. IEEE.
- [29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, Honolulu, HI, July 2017. IEEE.
- [30] Iris Ma, Alberto Krone Martins, and Cristina Videira Lopes. Integrating AI Tutors in a Programming Course, July 2024. arXiv:2407.15718 [cs].
- [31] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large Language Models: A Survey, February 2024. arXiv:2402.06196 [cs].
- [32] Rafael Padilla, Wesley L. Passos, Thadeu L. B. Dias, Sergio L. Netto, and Eduardo A. B. Da Silva. A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. *Electronics*, 10(3):279, January 2021.
- [33] Jeeseung Park. Jeeseung-Park/ViPLO, March 2025. original-date: 2023-04-12T11:41:18Z.
- [34] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. ViPLO: Vision Transformer based Pose-Conditioned Self-Loop Graph for Human-Object Interaction Detection, April 2023. arXiv:2304.08114 [cs].
- [35] Nitin Rane. YOLO and Faster R-CNN object detection for smart Industry 4.0 and Industry 5.0: applications, challenges, and opportunities. *SSRN Electronic Journal*, 2023.
- [36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, Las Vegas, NV, USA, June 2016. IEEE.
- [37] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. arXiv:1908.10084 [cs].
- [38] Zehua Sun, QiuHong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human Action Recognition from Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2022. arXiv:2012.11866 [cs].
- [39] Juan Terven and Diana Cordova-Esparza. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, November 2023. arXiv:2304.00501 [cs].
- [40] Janika Thielecke, Paul Couzy, and Gerben Hulsegege. Werkgevers Enquête Arbied 2024. 2024.
- [41] Yunjie Tian, Qixiang Ye, and David Doermann. YOLOv12: Attention-Centric Real-Time Object Detectors, February 2025. arXiv:2502.12524 [cs].
- [42] Liang Tianjiao and Bao Hong. A optimized YOLO method for object detection. In *2020 16th International Conference on Computational Intelligence and Security (CIS)*, pages 30–34, Guangxi, China, November 2020. IEEE.
- [43] Antonio van Dijck and Jesse van Schouten. CAKE: Context-Aware Knowledge Extraction Framework, February 2025.
- [44] Jesse J van Schouten. Moving Towards Tacit Knowledge Extraction, 2025.
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need, page 11, June 2017.
- [46] Yuxiao Wang, Qiwei Xiong, Yu Lei, Weiyi Xue, Qi Liu, and Zhenao Wei. A Review of Human-Object Interaction Detection, December 2024. arXiv:2408.10641 [cs].
- [47] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple Online and Realtime Tracking with a Deep Association Metric, March 2017. arXiv:1703.07402 [cs].
- [48] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, October 2024. arXiv:2303.18223 [cs].
- [49] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object Detection in 20 Years: A Survey, January 2023. arXiv:1905.05055 [cs].