



IDENTIFIKACE POPISŮ OBRÁZKŮ V TIŠTĚNÝCH DOKUMENTECH

2024/2025

Patrik Gáfrik (xgafri00)
Adrián Horváth (xhorva14)
Ondřej Bahounek (xbahou00)

1 Definice úlohy

Cílem je identifikovat části textu na stránce, které souvisejí s obrázky. Může jít o přímé popisky nebo textové pasáže, které obrázky popisují či vysvětlují.

Máme k dispozici rozsáhlý dataset fotografií stránek, z nichž část (≈ 6 tisíc) je označena štítky identifikujícími objekty jako obrázky, diagramy a popisky. Dále máme OCR výstupy těchto stránek, které obsahují text.

Rozhodli jsme se tuto úlohu rozdělit na dva dílčí problémy:

1. **Detekce popisků** – Natrénujeme neuronovou síť pro rozpoznávání popisků obrázků. Vstupem bude obrázek stránky spolu s detekcemi ohraňujících rámečků (bounding boxů) pro obrázky a popisky pod nimi.
2. **Vyhledávání popisků v textu** – Pomocí modelu CLIP, který je již natrénovaný na propojení textu s obrázky, budeme v OCR výstupu vyhledávat texty související s obrázky. Vstupem bude text z OCR, výstupem identifikace popisků v textu.

2 Detekce popisků pomocí Ultralytics YOLO

Pro detekci popisků pod obrázky jsme zvolili model YOLO od Ultralytics, který je díky své rychlosti, přesnosti a spolehlivosti jedním z nejlepších dostupných nástrojů pro detekci objektů v obrázcích. Tento framework nám umožnuje snadno model fine-tunovat na vlastních datech a přizpůsobit ho tak přesně našim potřebám [3].

Díky efektivní detekci a klasifikaci bounding boxů je Ultralytics YOLO ideální pro lokalizaci malých popisků u obrázků na stránkách, což přesně odpovídá cíli našeho projektu. Proto jsme jej vybrali jako hlavní nástroj pro řešení detekce popisků.

2.1 Trénování modelu

Model YOLOv8 jsme použili s předtrénovanými váhami, což nám umožnilo rychle začít s tréninkem na našem vlastním datasetu. Detekovali jsme dvě třídy: „Popisek pod obrázkem“ a „Obrázek“. Dataset jsme rozdělili na trénovací, validační a testovací část, aby bylo možné sledovat nejen schopnost modelu učit se, ale i jeho schopnost generalizovat na nová data.

Před tréninkem bylo nutné data správně připravit a převést do formátu, který YOLO očekává - normalizované souřadnice bounding boxů a přiřazení

příslušných tříd. Použili jsme tedy dvě třídy odpovídající našim požadavkům na detekci.

Trénink modelu probíhal po dobu 40 epoch s použitím adaptivního optimalizéru na modelu YOLOv8m, který obsahuje přibližně 25,9 miliónů parametrů. Sledujeme klíčové metriky jako loss, precision, recall a mean Average Precision (mAP).

Trénink byl proveden na výpočetním clusteru MetaCentrum, který poskytl potřebný výpočetní výkon pro efektivní zpracování a optimalizaci modelu.

Po dokončení tréninku jsme model otestovali na samostatné testovací množině, která nebyla během tréninku použita. Výsledky potvrdily, že model dobře generalizuje a dokáže spolehlivě detektovat jak popisky, tak samotné obrázky i na nových, neznámých datech.

2.2 Vyhodnocení

Proces vývoje modelu nebyl lineární a vyžadoval několik iterací trénování a ladění metodiky. Původně jsme detekovali pouze textové popisky pod obrázky, avšak brzy jsme si uvědomili, že pro kompletní výsledky experimentů je nezbytné detektovat i samotné obrázky, ke kterým popisky patří. To vedlo k zásadní změně v přístupu a úpravě anotací i výstupních tříd modelu.

V dalších fázích jsme narazili na problém značné variability typů obrázků ve vstupních datech. V anotacích se vyskytovaly kategorie jako Obrázek, Fotografie, Schéma, Reklama atd. V první iteraci jsme se soustředili pouze na detekci třídy Obrázek, ale později jsme rozšířili trénink i o třídu Fotografie, která často bývá opatřena popiskem stejně jako Obrázek. Naopak třída Reklama se v datech objevovala velmi často, ale téměř nikdy neobsahovala popisek. Z tohoto důvodu jsme se rozhodli ji z detekce vyloučit, aby model nebyl rušen nerelevantními vzory. Pro upřesnění, my tedy jako Obrázek detekujeme i objekty, které v původních anotacích mají třídu Fotografie.

V poslední řadě musíme spojit obrázky s jejich popisky, to je především důležité, pokud se na stránce nachází více obrázků a popisků. Ukázalo se, že spojení popisku k nejbližšímu obrázku funguje výborně, na podmniožině stránek, kde jsme to testovali, se nám podařilo se sto procentní úspěšností spojit obrázky s příslušným popiskem.

2.3 Výsledky

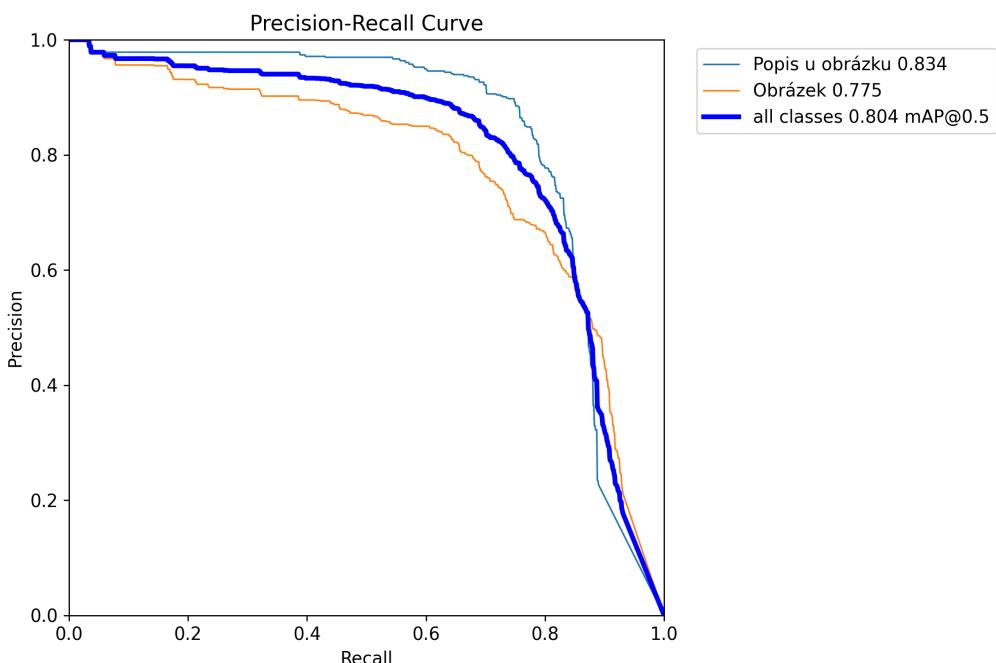
Finální model dosahuje velmi solidní přesnosti při detekci jak obrázků, tak popisků. Ve většině případů se mu daří správně identifikovat dvojice obra-

zových objektů a jejich textových anotací, a to i ve složitějších dokumentech s různorodým rozvržením.

Určitou výzvou zůstávají případy, kdy se v dokumentu nacházejí objekty jako reklama nebo schéma, které jsou vizuálně podobné třídě „obrázek“, ale zpravidla neobsahují žádný popisek. Tyto případy mohou vést k falešně pozitivním detekcím nebo ke snížení přesnosti klasifikace tříd. Přesto však model ve většině situací generalizuje velmi dobře a jeho výkon je pro potřeby našeho projektu plně dostačující.

Pro vyhodnocení výkonu jsme se zaměřili především na **Precision-Recall (PR) křivku**, která nejlépe vystihuje rovnováhu mezi úplností a přesností v našem detekčním úkolu. Vizuálně ilustruje, že model si vede stabilně napříč různými prahy detekce. Hodnota **mAP@0.5** pro obě třídy činí **0,804**, přičemž pro třídu **Obrázek** je to **0,775** a pro třídu **Popisek** **0,834**.

2.4 Výsledky



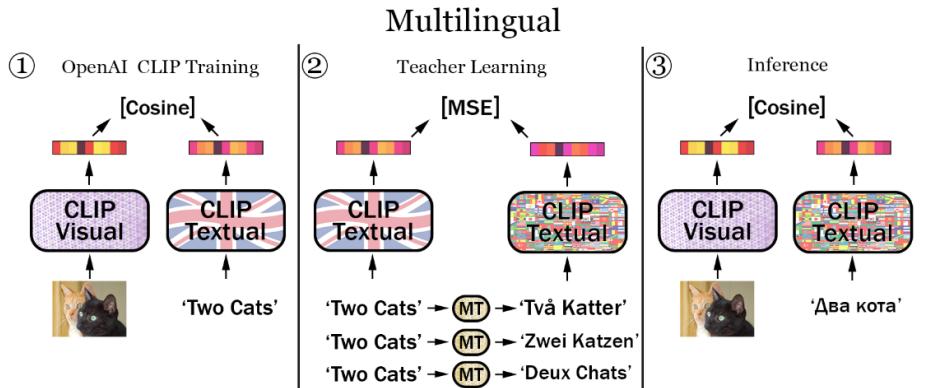
Obr. 1: Precision-Recall křivka finálního modelu YOLOv8m na verifikační datové sadě

3 Detekce popisků v textu pomocí modelu CLIP

Kromě detekce popisků u obrázků jsme se zaměřili také na identifikaci *popisků v textu*, které se mohou nacházet kdekoliv na stránce a nejsou přímo pod obrázky. Pro tento úkol jsme zvolili multimodální model CLIP (Contrastive Language–Image Pretraining) od OpenAI, který umožňuje porovnávat textové a obrazové vstupy v jednotném embedding prostoru [2].

Po testování jsme však dospěli k závěru, že tento model není pro náš případ dostačující, zejména kvůli nedostatečné podpoře českého jazyka. Proto jsme se rozhodli použít variantu Multilingual-CLIP, která je navržena tak, aby lépe pracovala s různými jazyky, nejen s angličtinou.

Multilingual-CLIP využívá nový textový transformer, který se při trénování porovnává s původní textovou částí modelu CLIP. Do nového modelu se vkládají popisky v různých jazycích, zatímco do původního modelu anglické ekvivalenty. Výstupní embeddingy obou modelů se porovnávají pomocí MSE ztráty, na jejímž základě se nový vícejazyčný model dále trénuje [1]. Takto natrénovaný textový model je pak možné kombinovat s původní obrazovou částí CLIPu pro vícejazyčné úlohy.



Obr. 2: Multilingual CLIP [1]

3.1 Postup detekce pomocí CLIP

Na vstupu máme obrázky jednotlivých stránek, ke kterým získáme bounding boxy všech obrázků – bud’ přímo z anotací datasetu, nebo z predikcí našeho YOLO modelu trénovaného na detekci obrázků a popisků. Současně máme k dispozici OCR výstupy s pozicemi a obsahem všech textových řádků na stránce.

Následně jsme provedli několik kroků předzpracování dat:

- **Odstranění popisků u obrázků:** Abychom se vyhnuli záměně popisků pod obrázky s popisky v textu, bylo nutné tyto texty z OCR výstupu vyfiltrovat. Pro každý detekovaný (nebo anotovaný) popisek pod obrázkem jsme porovnali jeho bounding box s bounding boxy OCR rádků a odstranili ty, které měly vysokou překryvnost (IoU).
- **Redukce vstupního textu:** Z důvodu omezení modelu CLIP, který zvládne maximálně 70 tokenů textu, jsme upustili od komplexních metod jako je vkládání celých textových bloků. Místo toho jsme zvolili přímý přístup - každý OCR rádek jsme vyhodnotili samostatně.

3.2 Vyhledávání popisků

Pro každý detekovaný obrázek jsme spočítali embedding pomocí CLIP. Následně jsme pro každý OCR rádek spočítali embedding textu a vypočetli kosinovou podobnost s embeddingem obrázku. Pokud byla podobnost vyšší než zvolený práh, označili jsme rádek jako potenciální popisek.

Abychom z jednotlivých rádků vytvořili smysluplné textové bloky, vybrali jsme nejlepší kandidáty (např. top 3 na stránku) a k nim jsme se pokusili přidávat předcházející i následující řádky — pokud také překračovaly práh podobnosti. Takto vzniklé bloky jsme považovali za finální textový popisek obrázku.

3.3 Vyhodnocení metody

Vyhodnocení detekce *popisků v textu* bylo výrazně náročnější než v případě popisků pod obrázky. Hlavním problémem je, že samotný pojem ”popisek v textu“ je značně subjektivní. Někteří by za popisek považovali pouze přímý komentář k obrázku, jiní zase jakýkoliv text, který popisuje objekt, jež se na obrázku nachází. Stejně tak není jednoznačné, jaký rozsah textu by měl být označen - má jít o jednu větu, celý odstavec, nebo pouze klíčová slova?

Navíc anotace v našem datasetu nejsou konzistentní a často ani přesné. Někdy anotace označují jen jedno slovo, jindy celou větu, odstavec nebo dokonce celou stránku. A počet popisků je v anotacích také sporný, někdy označí každé jméno předmětu na obrázku, někdy označí první zmínku o předmětu a někdy zmínku o předmětu nepovažují za popisek vůbec. Z tohoto důvodu není možné použít běžné automatizované metriky přesnosti nebo recallu jako spolehlivé měřítko.

Namísto toho jsme se při vyhodnocení zaměřili na to, zda naše metoda dokáže detekovat podobná místa jako anotace, případně zda zachytí texty,

které skutečně popisují obrázek - bez ohledu na to, zda přesně odpovídají označeným anotacím. Většina hodnocení tak probíhala manuálně: zkoumali jsme detekované pasáže a ověřovali, zda skutečně tematicky odpovídají zobrazeným obrázkům.

Přestože tedy nemůžeme poskytnout jednoznačné kvantitativní výsledky jako u předchozí metody, kvalitativní analýza ukazuje, že přístup s využitím CLIP často spolehlivě identifikuje relevantní texty související s obrázky.

Přesto pro jednodušší vyhodnocování parametrů modelu, jsme se dívaly i na pár metrik. Testovali jsme na náhodné submnožině 100 stránek, které obsahovaly popis v textu. Pro finální model jsme získali výsledky shrnuté v tabulce 1, která uvádí počet detekcí a intersekcí boxů detekcí s anotacemi.

Popis	Hodnota
Stránek celkem	100
Detekovaných popisků v textu	109
Počet popisků v textu dle anotací	49
Intersekcí celkem	23
Detekcí bylo intersekciemi	21 %
Stránek s aspoň jednou intersekcí	17 %

Tabuľka 1: Základní statistiky modelu

Je opět nutné připomenout, že jsme se nesnažili přesně detektovat stejné popisky jako v anotacích. Nicméně si myslíme, že počty intersekcí hovorí o tom, že odhadujeme zhruba podobné pasáže jako popisky.

Existují však i případy, kdy CLIP nefunguje příliš dobře. Některé obrázky zobrazují velmi staré předměty, jejichž pojmenování je zastaralé nebo dnes téměř nepoužívané. Přestože CLIP zvládá češtinu, mnoho technických vynálezů z devatenáctého století (které se v našem datasetu často objevují) nezná - ani názvem, ani vizuálně. To může vést k nesprávnému vyhodnocení podobnosti mezi obrazem a textem.

3.4 Vyhodnocení úlohy

Po vyhodnocení obou metod samostatně jsme je následně zkombinovali, aby chom zhodnotili fungování celého systému jako celku. Vzhledem k absenci přesných anotací a jasné definici detekovaných objektů jsme nepoužili žádné specifické metriky pro kvantitativní vyhodnocení.

Zaměřili jsme se proto na praktické výsledky a ověřili kvalitu kombinovaného výstupu manuální kontrolou na reálných stránkách. Hodnotili jsme,

zda systém detekuje obrázky, jejich přímé popisky a zároveň identifikuje části textu, které s obrázky významově souvisejí.

Výsledky ukazují, že náš přístup je schopen zachytit vztahy mezi vizuálním a textovým obsahem. Ve většině případů se model chová očekávatelně a dosažené výsledky dobře odpovídají lidskému chápání toho, co představuje popisek k obrázku.

Pro demonstraci přikládáme několik konkrétních příkladů výstupů, kde je vizualizováno spojení detekovaných obrázků, popisků pod nimi a odpovídajícího textu v hlavním obsahu stránky. (Při vizualizaci bloku textu, který detekujeme jako popisek v textu, kreslíme bounding boxy kolem každého řádku, i přestože ho detekujeme jako jeden blok, jak je vidět v obrázku 3)

Zdrojový kód je dostupný na GitHubu.¹ Natrénovaný YOLO model je ke stažení zde.²

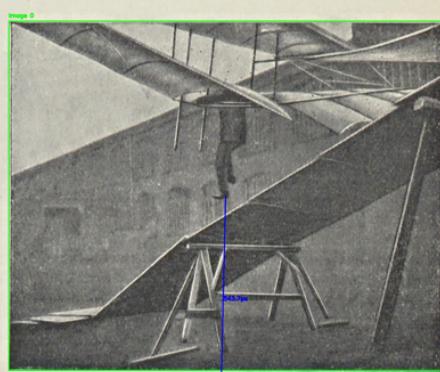
¹<https://github.com/PoweredByAdrian/KNN/tree/main>

²<https://drive.google.com/drive/folders/1D45xfXD6z0QNdKiQ8Ca0T7HDMqZWgzi>

tolik odborné zkušenosti, že by prý neměl do takového letadla vsednout nikdo, kdo si dříve nezískal náležitě obratnosti ve snásecím letu a neosvojil si instinctivní potřebnost k udržování rovnováhy, na níž ovšem závisí jeho život. Není prý daleka doba, kdy bude na školách v tělocviku závazně zavedeno cvičení ve snásecím letu.



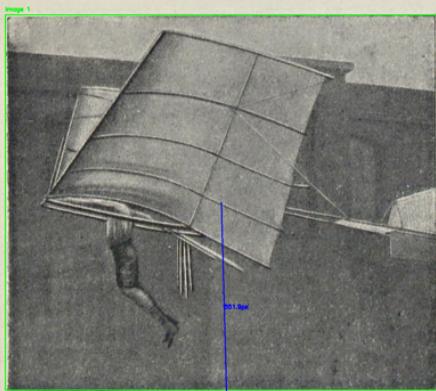
Obr. 255. První český aviatik inženýr
J. Kašpar.



Obr. 256. Vzlet snášecího letadla rozběhem po
nakloněné rovině.

Z časopisu: „Vynálezy a pokroky“.

Letadla, jichž jmenovaní průkopníci praktické aviatiky používali, byla lehká, tak že je aviatik snadno unesl na ramenech. Aby vzletl, musil se rozběhnout s letadlem se stráně proti větru, a když součet jeho vlastní rychlosti s rychlosí větru dosáhl určité velikosti, vneslo se letadlo s aviatickem od země a zvolna, klidně klesalo. Častým cvikem dospěli aviaticové tak daleko, že uletěli dráhu až několika set metrů. Podobně musí si počinat i dnes každý, kdo chce podniknout let se snášecím letadlem. Za účelem rozběhu postaví se vysoké dře-



Obr. 257. Počátek letu.



Obr. 258. Za letu.

Z časopisu: „Vynálezy a pokroky“.

věné lešení, s něhož vede k zemi nakloněná plocha, zhotovená z prken (obr. 256.). V zimě spojuje se aviatický sport s rohačkovým, za kterýmžto účelem upevňují se na rohačkách snášecí letadla. Jinak, totiž bez rozběhu po nakloněné ploše, startují členové vynikajícího francouzského aviatického spolku „Nord Aviation“ v Lille,

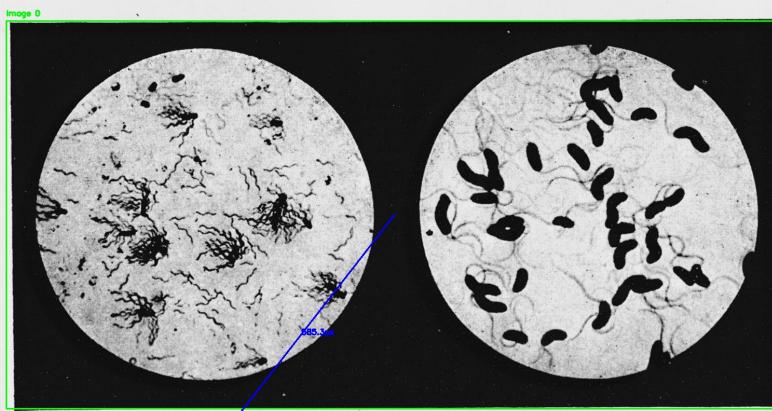
Okenní skla a lidské zdraví.

Ve schůzi rakouské společnosti pro péči o zdraví přednášel na jaře známý fysiolog dr. Hausmann o svých významných pokusech, na jejichž základě zjistil, že okenní skla, jichž se dnes užívá, jsou škodlivá našemu zdraví. Prof. Hausmann líčil vliv světla na lidský organismus a jeho význam pro jednotlivé životní procesy. Zejména ultrafialové paprsky působí profylakticky a léčivě na tuberkulosu a rachitis. Jest však již dávno známo, že okenní sklo těchto paprsků nepropouští. V Anglii na příklad užívá se proto pro nemocnice, školy atd. tak zv. okenních tabulí Vita, které propouštějí ultrafialové paprsky, obsažené v slunečním světle, ale jsou příliš drahé a proto se jich poměrně málo užívá. Prof. Hausmann zjistil četnými pokusy, že i různé levné druhy skla propouštějí ultrafialové paprsky, jež jsou tak důležité pro lidské tělo. Přednášející požádal, aby byl brán na tuto okolnost patřičný zřeteł v novostavbách a poznamenal, že v Japonsku nebyla tak dlouho rachitis, dokud se tam nezačalo stavět po evropském způsobu. S okenním sklem přišla do Japonska i rachitis (křivice). (Sklářské rozhledy.)

Pohyb bakterií

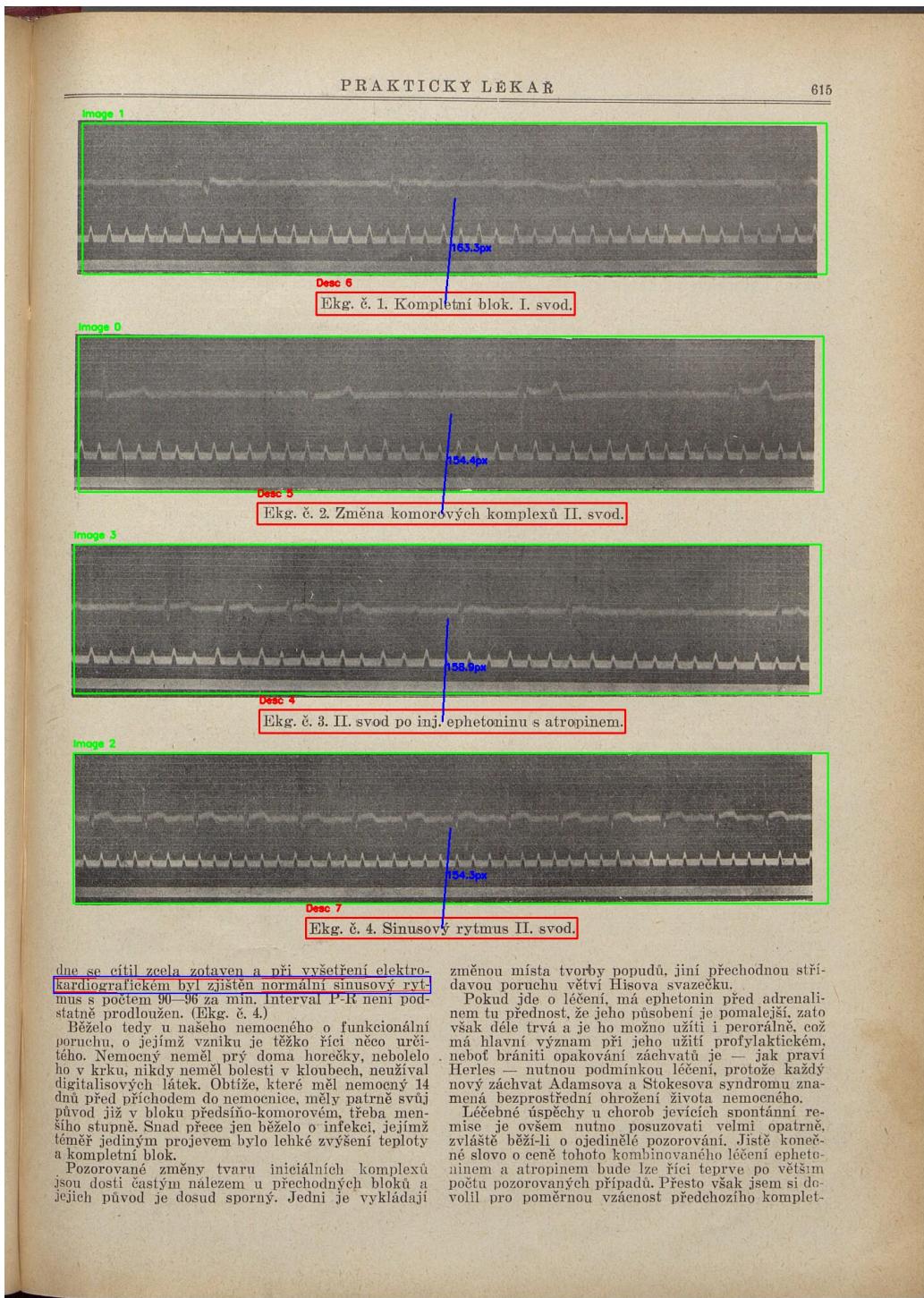
je podmíněn jemnými řasinkami, jimiž bakterie víří, jsou-li v půdě, která dovoluje pohybu (ve vodě, ve výkalech, v krvi).

Tyto řasinky se velmi nesnadno dokazují pod drobnohledem a nutno užít zvláštních barvicích metod, aby se staly zjevnými našemu oku. Zde ukázka dvou druhů bakterií, obdařených řasinkami.



Rovněž zárodky tyfu střevního mají řasinky.

Obr. 4: Příklad detekce 2



dnu se cítil zcela zotaven a při vyšetření elektro-kardiografickém byl zjištěn normální sinusový rytmus s počtem 90–96 za min. Interval P-R není podstatně prodloužen. (Ekg. č. 4.)

Běželo tedy u našeho nemocného o funkcionální poruchu, o jejímž vzniku je těžko říci něco určitého. Nemocný neměl prý doma horčeky, nebolelo ho v krku, nikdy neměl bolesti v kloboucích, neužíval digitalisových láték. Obtíže, které měl nemocný 14 dnů před příchodem do nemocnice, měly patrně svůj původ již v bloku předsínno-komorovém, třeba menší stupně. Snad přece jen běželo o infekci, jejímž témař jediným projevem bylo lehké zvýšení teploty a kompletní blok.

Pozorované změny tvaru iniciačních komplexů jsou dosť častým nalezem u přechodných bloků a jejich původ je dosud sporný. Jedni je vykládají

změnou místa tvorby popudu, jiní přechodnou střídavou poruchu větví Hisova svazečku.

Pokud jde o léčení, má ephetonin před adrenalinem tu přednost, že jeho příslušení je pomalejší, zato však déle trvá a je ho možno užít i perorálně, což má hlavní význam při jeho užití profylaktickém, neboť bránit opakování záchvatů je — jak praví Herles — nutnou podmínkou léčení, protože každý nový záchvat Adamsova a Stokesova syndromu znamená bezprostřední ohrožení života nemocného.

Léčebné úspěchy u chorob jevících spontánní remisie je ovšem nutno posuzovat velmi opatrně, zvláště běží-li o ojedinělé pozorování. Jistě konečné slovo o ceně, tohoto kombinovaného léčení ephetoninem a atropinem bude lze říci teprve po větším počtu pozorovaných případů. Přesto však jsem si dovolil pro poměrnou vzácnost předchozího komple-

Obr. 5: Příklad detekce 3

Referencie

- [1] Fredrik Carlsson. *Multilingual-CLIP: OpenAI CLIP text encoders for multiple languages.* <https://github.com/FreddeFrallan/Multilingual-CLIP>. Accessed: 2025-05-15. 2022.
- [2] Alec Radford, Jong Wook Kim, Luke Hallacy et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *arXiv preprint arXiv:2103.00020* (2021).
- [3] Ultralytics. *Ultralytics YOLO Documentation.* <https://docs.ultralytics.com/>. Accessed: 2025-05-15. 2025.