Cíl EÚS
Česká republika – Svobodný stát Bavorsko
2014–2020

Přírodovědecká fakulta
Faculty of Science
Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Evropská unie
Evropský fond pro regionální rozvoj

Methodik der Entwicklung von Software-Tools für die automatische Erkennung von Sakralgebäuden

Jan Fesl , Jiří Jelínek, Michal Konopa, Marie Feslová a Kateřina Horníčková.

ABSTRAKT

Die zuverlässige Identifizierung bestimmter Orte, Objekte, Personen usw. anhand von Fotografien ist eine in der Praxis sehr häufige Aufgabe, die nicht einfach ist und viele Herausforderungen mit sich bringt. Bei einer großen Anzahl von Fotos ist die manuelle Identifizierung der darauf abgebildeten Objekte aus Zeitgründen fast unmöglich und muss durch Softwaresysteme automatisiert werden. In der Praxis stellt sich heraus, dass es sehr schwierig ist, ein einziges universelles Identifikationssystem zu schaffen, das eine bestimmte Person, ein bestimmtes Tier oder ein bestimmtes Gebäude auf Fotos von z. B. Menschen, Tieren oder Gebäuden ohne zusätzliches Expertenwissen identifizieren kann. Noch schwieriger ist die Situation, wenn die Anzahl der Fotos, die als Referenz dienen, sehr gering ist. Die Methoden, die zur Objektidentifizierung eingesetzt werden können, sind zwar allgemeingültig, ihre Anwendung muss jedoch an das jeweilige Szenario angepasst werden. Unser Forschungsteam hat ein Softwaresystem entwickelt, das mit sehr gutem Erfolg ein bestimmtes Objekt auf Fotos von eng umrissenen Objekten (Sakralbauten - Kirchen oder Kapellen) bzw. deren Innen- oder Außenbereich zuverlässig identifizieren kann. Der von uns gewählte Ansatz scheint für das von uns gewählte Gebiet vielversprechend zu sein und lässt sich leicht auf andere Bereiche übertragen.

SCHLÜSSELWÖRTER

Sacral objects, identification, machine learning, Siamese networks, neural networks, expert system.

1. EINFÜHRUNG UND MOTIVATION

Im Rahmen unseres Projekts "Informationssystem für mittelalterliche Denkmäler" haben wir an einer Forschungsaufgabe gearbeitet, die es Nutzern erleichtern soll, bestimmte auf Fotos abgebildete Orte zu identifizieren, indem sie diese mit einer Bilddatenbank abgleichen. Eine grundlegende Annahme, die diese Aufgabe nicht einfach macht, ist, dass es in der Praxis eine sehr kleine Anzahl von Bildmustern für bestimmte Objekte gibt und daher klassische Bildverarbeitungsmethoden, die auf tiefen neuronalen Netzen basieren, nicht verwendet werden können, da diese eine große Anzahl von Beispielen benötigen. Eine weitere Annahme, die die Identifizierungsaufgabe erschwert, besteht darin, dass keine zwei Fotos in der Bilddatenbank genau mit dem zu analysierenden Foto übereinstimmen.

Die Inspiration für die Lösung dieses Problems ist, dass eine Person manchmal anhand eines einzigen Referenzfotos, das sie mit anderen Fotos vergleicht, feststellen kann, ob es sich um

identische Objekte handelt. Im Allgemeinen erfolgt die Identifizierung nie auf der Grundlage des Ganzen, sondern nur anhand ausgewählter spezifischer Details. Diese Annahme haben wir bei der Entwicklung unseres Systems zugrunde gelegt.

2. THEORETISCHE EINFÜHRUNG

Die Schaffung eines Systems zur Identifizierung sakraler Objekte erfordert die Zusammenarbeit von Experten aus mehreren Bereichen, nämlich: Geschichte, Kunst- und Kulturgeschichte, Architektur (zusammen Kunst und Geschichte) und fortgeschrittene Methoden des maschinellen Lernens oder der Computer Vision (zusammen künstliche Intelligenz - KI). Experten aus dem Bereich Kunst und Geschichte wissen, welche spezifischen Komponenten (wie Altäre, Presbyterien oder Baptisterien) in Fotografien sakraler Objekte einzigartig sind und können daher zur eindeutigen Identifizierung eines bestimmten Objekts beitragen. Experten auf dem Gebiet der Künstlichen Intelligenz kennen Methoden, die automatisch bestimmte Objekte in Fotografien finden können, und andere Methoden, die die gefundenen Komponenten mit Mustern in einer Bilddatenbank abgleichen können.

2.1 Die Meinung eines Experten auf dem Gebiet der Kunstgeschichte

Der Wechsel der künstlerischen Stile kennzeichnet die Entwicklung der Kunst. Ein Stil ist ein kohärenter künstlerischer Ausdruck einer Epoche, der sich durch eine kontinuierliche Kombination spezifischer Formen und Merkmale auszeichnet, die für eine bestimmte Zeit typisch sind und im Werk zeitgenössischer Künstler Anwendung finden [4] [5]. Ein Stil stellt eine unterteilte und von außen erkennbare Menge visueller Zeichen dar, mit denen ein Objekt geschaffen wird, und bildet die formale Sprache eines Werks, ähnlich wie Grapheme im geschriebenen Text [5]. Wir erforschen diese durch formale Analyse. Einige dieser Elemente sind leicht zu erkennen, während andere schwieriger zu identifizieren sind, da sie aus dem gleichzeitigen Auftreten und der Interaktion mehrerer spezifischer Elemente bestehen [6].

In der Architektur manifestieren sich Standardelemente sowohl in der konventionellen Raumgestaltung der jeweiligen Epoche als auch in den formalen Details des Gebäudes, die diese kennzeichnen. Die Sammlung dieser Elemente hilft, die strukturelle Entwicklung eines bestimmten Gebäudes im Laufe der Geschichte aufzuzeigen. Die Beobachtung und Bestimmung des Sachverständigen beruht auf der Ähnlichkeit und Identifizierung dieser charakteristischen Elemente und ihrer Kombinationen [8], die im Laufe der Ausbildung durch Beobachtung schrittweise erlernt werden.

Die Kirchen in der tschechischen Umgebung haben eine komplexe architektonische Entwicklung und viele Veränderungen durchlaufen. Die daraus resultierende Struktur ist oft eine Mischung aus Details aus verschiedenen Epochen; mittelalterliche (romanische und gotische), die die Grundlage der untersuchten Gebäude in der südböhmischen Region bilden, charakteristische gebündelte romanische Fenster, längliche gotische Fenster, gotische polygonale Endstücke, gebrochene gotische Triumphbögen oder Heiligtümer, Sanctus-Türme, quadratische Türme, usw.

Im Mittelalter und in der frühen Neuzeit gab es keine Massenproduktion. Die Baufabriken schufen für den Ort einzigartige Elemente, die der Morphologie des damaligen Stils entsprachen, aber auch den Erfindungsreichtum der Steinmetze einbrachten [9]. Obwohl jedes Detail des Gebäudes individuell ist, da es als Einzelstück geschaffen wurde, weist es allgemeine Merkmale in einer für

die mitteleuropäische Architektur und die Region typischen Form auf. So lässt sich ein Element nicht nur einer Epoche und einer geografischen Region oder einer Reihe von Bauten, sondern auch diesem speziellen Bauwerk nach seinen einzigartigen Merkmalen und seiner Anordnung zuordnen (Wölfflin, 1915, 149-150) [4].

2.2 Aktuelle Trends in der Computer Visionní

Wie bereits erwähnt, basiert die Datenverarbeitung auf Verfahren zur Erkennung von Objekten im Bild. Diese Verfahren lassen sich grob in zwei Gruppen einteilen: einstufige und mehrstufige Verfahren. Bei mehrstufigen Verfahren werden die Aufgaben der Identifizierung potenzieller objektdefinierender Regionen (Regionen), der Auswahl von Regionen für die weitere Verarbeitung und der Identifizierung der Objektkategorie innerhalb einer gegebenen Region getrennt behandelt. Typische Modelle mit mehrstufiger Verarbeitung sind zum Beispiel Netze wie RPN, RCNN, FastRCNN, FasterCNN oder MaskCNN. Modelle aus dieser Gruppe zeichnen sich dann durch die hohe Qualität der durchgeführten Identifikation aus, aber leider auch durch die geringere Verarbeitungsgeschwindigkeit der Eingangsdaten, was ihren Einsatz in Echtzeit einschränkt.

Die Gruppe der Modelle mit einstufiger Verarbeitung umfasst hauptsächlich Modelle des Typs YOLO. Diese Modelle lösen die oben genannten Einzelschritte innerhalb der komplexen Verarbeitung. Diese Modelle sind wesentlich schneller und können auch für die Online-Objekterkennung verwendet werden. Andererseits sind sie durch eine etwas geringere Erkennungsqualität und einige andere Probleme (z. B. Schwierigkeiten bei der Erkennung kleiner Objekte) gekennzeichnet.

Für die erste Phase der Verarbeitung der Eingangsbilddaten wurde für dieses Projekt das YOLO-Modell gewählt, und zwar die Version YOLOv5, die zum Zeitpunkt der Antragserstellung die letzte der Entwicklungsreihe war. Gegenwärtig sind auch die Bezeichnungen YOLO6 und YOLO7 zu sehen. Die Version 6 verwendet die gleiche Schnittstellenmethode wie die Version 5, so dass es möglich ist, dass Skripte des Modells der Version 5 mit Gewichten verwendet werden, die als Gewichte der Version 6 bezeichnet werden, wobei die Hauptverbesserung hier in der höheren Auflösung des Eingabebildes besteht. Bei der Version 7 herrscht eine gewisse Verwirrung über den Namen (es gibt zwei konkurrierende und inkompatible Versionen), und auch die Struktur und die Schnittstelle des Modells sind anders. Es wird jedoch davon ausgegangen, dass diese Version wesentliche Verbesserungen mit sich bringt, insbesondere die Fähigkeit, das identifizierte Objekt genau abzugrenzen (was für unsere Anwendung nicht notwendig ist).

Die Bezeichnung Yolov5 bezieht sich jedoch nicht auf ein einzelnes Modell, sondern auf eine ganze Gruppe von Modellen, die sich in Struktur, Art und Anzahl der verwendeten Schichten innerhalb des Modells unterscheiden..

YOLO-Modelle können auf klassische Weise trainiert werden, d. h. von Anfang an mit Zufallsgewichten. Dieser Prozess ist jedoch sehr langsam und es ist besser, die Technik des Transfer-Lernens anzuwenden und das Modell mit bereits trainierten Gewichten zu verwenden. Es sind Gewichte verfügbar, die mit dem COCO-Datensatz trainiert wurden. Diese wurden in unserem Projekt verwendet, aber es war notwendig, den gewählten Modelltyp neu zu trainieren. Der Grund dafür war, dass unser Projekt mit bestimmten sakralen Objekten arbeitet, die im COCO-Datensatz nur sehr schlecht vertreten sind. Es wurde auch erwogen, einen anderen

standardisierten Datensatz zu verwenden, der zur Verfügung stehen würde. Die Autoren haben jedoch keinen Datensatz gefunden, der für die Klassifizierung dieser Art von Objekten geeignet ist.

Daher wurde der bereits erwähnte Datensatz zum Trainieren der Modelle verwendet..

Die erste Phase der Datenverarbeitung für historische Denkmäler wurde oben beschrieben. Sie basiert auf der Identifizierung von Objektkategorien, die in den Eingabebildern erkannt wurden. Das Ergebnis ist also eine Liste der auf dem Bild identifizierten Objekte, einschließlich ihrer Klassifizierung und ihres Standorts, so dass für sie entsprechende Ausschnitte aus dem Originalbild erstellt werden können. Diese Ausschnitte und die Klassifizierungsinformationen dienen dann als Input für die weitere Verarbeitung.

Die zweite Phase der Datenverarbeitung in diesem Projekt schließt sich an die soeben beschriebene erste Phase an, d. h. die Erkennung von Objekten im Bild und ihre Kategorisierung. Um die Ziele des Projekts zu erreichen, muss das Objekt nicht nur kategorisiert werden, sondern es muss auch versucht werden, es spezifisch als eine Instanz innerhalb der Kategorie zu identifizieren.

In Phase 2 wird also versucht, innerhalb einer bestimmten Kategorie festzustellen, ob zwei gegebene Bilder das gleiche Objekt abbilden. Die grundlegende Aufgabe besteht darin, die Ähnlichkeit zweier Eingabebilder zu untersuchen, die Objekte desselben Typs abbilden. Für diese Aufgabe gibt es eine Reihe von Techniken, die sich mit der Bildanalyse befassen, sowohl spezielle als auch andere, die ursprünglich für andere Arten von Aufgaben entwickelt wurden. Diese Techniken basierten ursprünglich auf klassischen Bilddatenverarbeitungstechniken und verwendeten keine auf maschinellem Lernen basierenden Techniken. Diese Techniken lassen sich in zwei grundlegende Gruppen einteilen: flächenbasierte und merkmalsbasierte.

Bei flächenbasierten Techniken ist die Analyse des Bildes in seinen einzelnen Teilen oder Pixeln die Grundlage. Hier kann die Ähnlichkeit von Bildern z. B. auf Informationen über die Farbe von Teilen der Bilder beruhen (Vergleich von Farbhistogrammen). Erinnern wir uns daran, dass wir in der ersten Phase nur Ausschnitte haben, die die identifizierten Objekte erfassen, und dass die obige Technik auch aus diesem Grund erfolgreich sein kann. Sie reagiert jedoch sehr empfindlich auf Unterschiede in der Helligkeit der Bilder und auf Unterschiede in der Farbwiedergabe.

Merkmalsbasierte Techniken konzentrieren sich auf die Extraktion von Merkmalen aus dem Bild und deren anschließende Verarbeitung. Eine der Methoden besteht darin, Schlüsselpunkte im Bild in Graustufen (eindeutig identifizierbar) zu identifizieren und dann die gefundenen Punkte in beiden Bildern hinsichtlich ihrer Lage und relativen Position zu vergleichen. Das Verfahren wird als skaleninvariante Merkmalstransformation (SIFT) bezeichnet und ist in der Lage, in der Ausgabe eine entsprechende Darstellung zu finden, die beschreibt, wie das eine Bild in das andere transformiert werden kann. Sie ist in erster Linie für Anwendungen gedacht, die genau diese Transformation erfordern, z. B. die Fusion von Panoramabildern. Das Verfahren ist in der Lage, Ähnlichkeiten selbst zwischen Bildern zu erkennen, die sich in Bezug auf Helligkeit, Position und Fokus des Betrachters, Drehung, Zoom usw. erheblich unterscheiden. Ihr Einsatz ist eher hardware-intensiv..

Eine andere Technik, die so genannten beschleunigten robusten Merkmale (SURF), basiert auf SIFT und enthält dieselben globalen Schritte, unterscheidet sich aber in der konkreten Umsetzung von SIFT. Die Verwendung von SURF ist weniger ressourcenintensiv und die Berechnung ist daher schneller.

Die oben genannten Techniken verwenden keine Prinzipien des maschinellen Lernens. Dies hat jedoch völlig neue Impulse und Techniken in den Bereich der Grafikverarbeitung gebracht. Der Schwerpunkt liegt hier auf dem Lernen mit dem Lernenden, was eine bessere Anpassung des Modells an eine bestimmte Aufgabe ermöglicht. Allerdings ist eine ausreichend große und mit Anmerkungen versehene Trainingsmenge erforderlich.

Siamesische Netze scheinen eine geeignete Architektur für den Abgleich von erkannten Objekten mit einer Referenzdatenbank zu sein, da sie zwei abzugleichende Bilder als Eingabe haben und in der Lage sind, ein Maß für ihre gegenseitige Übereinstimmung als Ausgabe zu liefern. Diese Netze gehören zur Kategorie der merkmalsbasierten Werkzeuge. Auch hier ist es wichtig, daran zu denken, dass wir mit einer sehr spezifischen Gruppe von Objekten in Bildern arbeiten, nämlich mit sakralen Objekten. Daher ist es auch hier notwendig, das Modell mit den spezifischen Daten aus dem oben beschriebenen Datensatz vorzutrainieren

Das Netz hat zwei Eingänge (Bilder mit gleicher Auflösung und Farbtiefe), die zunächst von der ersten Stufe des Modells auf der Grundlage von Faltungsschichten verarbeitet werden, die die besten Ergebnisse bei der Verarbeitung grafischer Informationen und der Erfassung der in den Bildern auftretenden Grundmuster (Zeichen) erzielen. Diese Schichten werden durch einen Überbau aus mehreren dichten Schichten ergänzt, und diese gesamte Struktur bildet ein Untermodell, das im Folgenden als Turm bezeichnet wird. Der Turm wird mit gleicher Gewichtung für die Verarbeitung beider Eingangsbilder verwendet, seine Gewichte werden also geteilt. Auch hier kann das Turm-Submodell von Grund auf neu erstellt und nur mit den von uns gelieferten Eingabedaten trainiert werden. Es können aber auch bereits bekannte Verfahren und Strukturen verwendet werden, insbesondere für die Erstverarbeitung von Bilddaten mit CNN-Schichten. Hier kommen insbesondere die vortrainierten Modelle VGG-16, Inception (v3), ResNet50 oder EfficientNet in Betracht. Somit kann das Transfer-Lernen, also die Übernahme der Einstellungen dieser Modelle, die auf einem der bekannten vorhandenen Datensätze erstellt wurden, vorteilhaft genutzt werden. In der vorgestellten Lösung wurde das VGG16-Modell (ohne Aufbauschichten) mit den auf dem ImageNet-Datensatz erstellten Gewichtseinstellungen verwendet. Auf diese Struktur folgt dann eine mehrschichtige Struktur, die aus vollständig verbundenen Schichten besteht. Im Vergleich zur Eingabe, die z. B. für VGG16 eine Dimension von 224 x 224 x 3 hat, ergibt sich eine deutliche Dimensionsreduktion der Ausgabe.

Das Ergebnis der Turmstufe sind zwei Vektoren der gleichen Dimension, einer für jedes Bild. Diese Vektoren bilden den Input für die zweite Stufe des Modells, die dazu dient, die Bilder zu vergleichen und ihren Ähnlichkeitsgrad zu bestimmen. Das Teilmodell besteht aus nur wenigen Schichten. Die Rolle der ersten Schicht ist entscheidend, da sie den euklidischen Abstand zwischen den eingereichten Vektoren berechnet. Die nachfolgenden Schichten normalisieren diesen Wert dann lediglich auf den Bereich (0, 1) und wandeln ihn in eine einzige Ausgabe um, die das Ähnlichkeitsmaß der vorgelegten Bilder angibt. Die Einstellung der Verlustfunktion des gesamten Modells ist ebenfalls einer der Schlüsselfaktoren für den Erfolg des Modells. In dem vorgestellten Modell verwenden wir eine Verlustfunktion aus (https://keras.io/examples/vision/siamese_contrastive/), die dazu dient, die Unterschiede zwischen den Bildern hervorzuheben.

3. METHODOLOGIE

Für das Training von Yolo- oder Siamesischen Neuronalen Netzen ist ein qualitativ hochwertiger Datensatz erforderlich, den wir in diesem Fall von Grund auf neu erstellen mussten. Auf der Grundlage der Empfehlungen von Experten wählten wir einige grundlegende Komponenten aus, die häufig in Sakralbauten zu finden sind und daher zur Identifizierung bestimmter Objekte verwendet werden können. Außerdem entwarfen wir ein zweistufiges System zur Objektidentifizierung, das wir in unserem System implementierten.

3.1 Erstellen eines Datensatzes

Da wir für unsere Zwecke 6000 Fotos von echten sakralen Gebäuden gemacht haben, haben wir sie manuell inspiziert und die relevanten Komponenten identifiziert. Die Gesamtgröße unseres Datensatzes betrug etwa 94 GB. Um eine bestimmte Klassifizierung von Bauteilen erstellen zu können, war es notwendig, eine spezielle Anwendung für diesen Zweck zu entwickeln, die wir PhotoCutter nannten. Eine kurze Beschreibung der Anwendung finden Sie weiter unten..

3.1.1 PhotoCutter

Die PhotoCutter-Anwendung dient dazu, Ausschnitte von gesuchten Objekten aus Fotos zu erstellen und diese dann in Dateien zu speichern. PhotoCutter ist eine Java-Anwendung mit einer grafischen Benutzeroberfläche, die auf der Java Swing-Bibliothek basiertg.

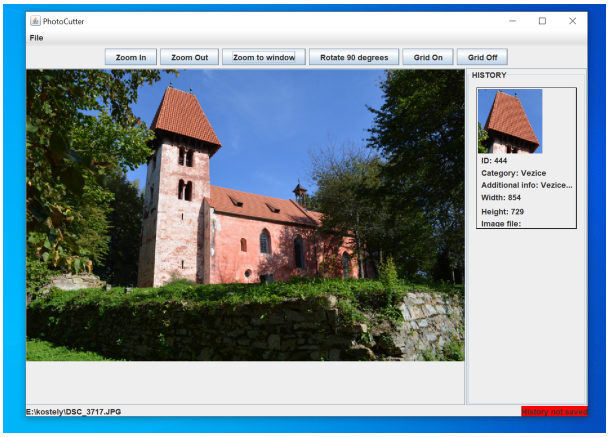

Bild. 1: PhotoCutter

In der nachstehenden Abbildung sehen Sie, wie die Klassifizierung der einzelnen Fotografien von sakralen Objekten oder die Suche nach den einzelnen in Tabelle 1 aufgeführten Bestandteilen durchgeführt wurde.

3.2 Struktur des Datensatzes

Testreihen haben wir auf der Grundlage unserer beruflichen Erfahrung einige wesentliche Details und Bauelemente ausgewählt, z. B. Statue, Fresko, Altar, Kanzel, Turm, Fenster, Kirchturm/Glockenturm, usw. Auch ihre Kombination ist wesentlich - nur ihre Kombination schafft eine eindeutige Identifizierung eines bestimmten Gebäudes und die Grundlage für die zeitliche, stilistische und formale Einordnung und Bewertung des Gebäudes sowie für die eigentliche Identifizierung des Gebäudes. Wandmalereien sind aufgrund ihrer Form leicht identifizierbar, weshalb auch zeitgenössische Kirchenmalereien zu den Elementen gezählt werden. Die Kirche kann anhand des stilistischen Niveaus und der Ikonographie der Innenmalereien identifiziert werden.u.

Eine vollständige Liste der Kategorien, einschließlich Beispielvisualisierungen, findet sich in Tabelle 1. Die Diagramme in der rechten Spalte stellen ein Histogramm der Bildverteilungen dar, sortiert nach Bildfläche, ausgedrückt in Kilopixeln (Kpix).

4. ERGEBNISSE UND MESSUNGEN

4.1 Erkennung von Objekten

Die zum Trainieren des Modells verwendeten Daten umfassten insgesamt 1097 Bilder mit 5164 gekennzeichneten Objekten. Die Testdaten bestanden aus 239 Bildern mit 1240 Objekten. Die Testdaten wurden in keiner Weise für den Lernprozess der verschiedenen Teile des Modells verwendet.

Die erste Phase der Datenverarbeitung im Modell wird durch das YOLOv5-Netz dargestellt. Wie bereits erwähnt, gibt es mehrere Netze dieses Typs, wobei die folgenden Varianten getestet wurden:

Model	Resolution
yolov5m	640
yolov5x	640
yolov5s6	1280
yolov5m6	1280
yolov5n	640
yolov5n	640

Die Modelle wurden für 400 Epochen auf dem Trainingsset trainiert. Anschließend wurden die angepassten Modelle anhand des Testsatzes bewertet. Die Werte der Schlüsselparameter, die sowohl für das Training als auch für das Testen charakteristisch sind, werden in der folgenden Tabelle zusammengefasst:

	Train	Test
Model	**P R mAP@.5 mAP@**	**P R mAP@.5 mAP@**
yolov5m	0.983 0.857 0.917 0.856	0.953 0.882 0.899 0.814
yolov5x	0.986 0.91 0.934 0.907	0.974 0.878 0.896 0.838
yolov5s6	0.978 0.865 0.926 0.869	0.949 0.877 0.9 0.812
yolov5m6	0.985 0.912 0.934 0.908	0.976 0.888 0.898 0.836
yolov5n	0.981 0.844 0.949 0.762	0.953 0.852 0.891 0.685
yolov5n	0.956 0.822 0.872 0.68	0.923 0.845 0.881 0.67

Das Modell "yolov5x" wurde auf der Grundlage des mAP@-Wertes, der bei der Auswertung der Testreihe erzielt wurde, für die weitere Bildverarbeitung ausgewählt. Das Modell bestand aus 444 Schichten und hatte insgesamt 86328181 Parameter.

Die folgende Tabelle zeigt die wichtigsten Daten aus dem Training des ausgewählten Modells nach Objektkategorien:

Class	Images	Labels	P	R	mAP@.5	mAP@
all	1097	5164	0.986	0.91	0.934	0.907
Detail ostění/přípory/konzoly	1097	389	1	0.992	0.995	0.945
Freska	1097	1352	0.999	0.996	0.995	0.978
Kaple	1097	36	0.993	1	0.995	0.986
Kazatelna	1097	41	0.994	1	0.995	0.985
Klenba	1097	190	0.999	1	0.995	0.989
Kostel	1097	241	1	0.994	0.995	0.993
Kruchta/Empora	1097	3	0.807	1	0.995	0.995
Křtitelnice	1097	19	0.978	1	0.995	0.964
Obraz	1097	120	0.995	1	0.995	0.97
Okno	1097	1393	0.999	0.991	0.995	0.95

Oltář	1097	129	0.998	1	0.995	0.987
Pilíř/Sloup	1097	339	0.995	0.997	0.995	0.961
Portál	1097	143	0.998	1	0.995	0.985
Průčelí	1097	1	0.965	1	0.995	0.995
Římsa	1097	4	1	0	0.25	0.15
Sanktuář/Pastoforium	1097	25	1	0.976	0.995	0.931
Socha	1097	141	1	1	0.995	0.955
Svorník	1097	210	1	0.999	0.995	0.915
Varhany	1097	19	0.987	1	0.995	0.98
Věžice/Zvonice/Sanktusník na střeše	1097	109	0.992	1	0.995	0.957
Vítězný oblouk	1097	7	0.981	1	0.995	0.975
Věž	1097	252	0.999	0.992	0.995	0.986
Závěr kostela/Presbytář	1097	1	1	0	0.332	0.332

Für jede Objektkategorie ist auch eine Testausgabe verfügbar:

Class	**Images**	**Labels**	**P**	**R**	**mAP@.5**	**mAP@**
all	239	1240	0.974	0.878	0.896	0.838
Detail ostění/přípory/konzoly	239	109	0.96	0.817	0.905	0.807
Freska	239	294	0.989	0.932	0.96	0.9
Kaple	239	3	0.983	1	0.995	0.897
Kazatelna	239	12	0.995	1	0.995	0.964
Klenba	239	39	1	0.815	0.895	0.813

Kostel	239	61	1	0.938	0.985	0.949
Kruchta/Empora	239	2	0.838	0.5	0.507	0.507
Křtitelnice	239	3	0.98	1	0.995	0.995
Obraz	239	21	0.898	0.857	0.851	0.826
Okno	239	313	0.993	0.915	0.983	0.888
Oltář	239	40	1	0.954	0.982	0.926
Pilíř/Sloup	239	109	0.97	0.898	0.921	0.865
Portál	239	32	0.997	1	0.995	0.949
Římsa	239	1	1	0	0	0
Sanktuář/Pastoforium	239	4	0.983	1	0.995	0.971
Socha	239	37	0.998	0.946	0.951	0.904
Svorník	239	67	0.956	0.967	0.943	0.772
Varhany	239	2	0.962	1	0.995	0.995
Věžice/Zvonice/Sanktusník na střeše	239	22	1	0.959	0.995	0.939
Vítězný oblouk	239	2	0.967	1	0.995	0.798
Věž	239	67	0.983	0.94	0.973	0.94

Die letzte Kategorie des Trainingssatzes ist in den Testdaten nicht vertreten und wird daher nicht aufgeführt.a.

Die unterschiedliche Qualität der Objekterkennung wird sowohl durch die Anzahl der Beispiele in jeder Kategorie als auch durch das vorherige Training des YOLO-Netzes beeinflusst, das nur auf unseren Daten neu trainiert wurde.

4.2 Identifizierung von Objekten anhand der Musterdatenbank

In der zweiten Phase wurde ein Siamesisches Netzmodell mit dem Faltungsunterteil VGG 16 verwendet, das auf dem Imagenet-Datensatz vortrainiert wurde und für Gewichtsänderungen gesperrt ist. Der Überbau bestand aus einer dichten Schicht sowie mehreren Schichten, um die Ausgaben der "Turm"-Teilmodelle für die beiden Eingabebilder zu kombinieren. Nach den ersten Tests wurde eine zweite Variante der Modellstruktur hinzugefügt, bei der die letzte Schicht VGG16 entfernt wurde, wodurch sich die Anzahl der einstellbaren Parameter (Modelle mit der Bezeichnung "strip") deutlich verringerte. Die Anzahl dieser Parameter stieg dann von 69.796 auf 3.215.524. Zwei Varianten der Verlustfunktionsbewertung, mse und Kontrastverlust (siehe oben), wurden ebenfalls getestet.

Die Eingabe-Trainingsdaten bestanden aus Objektausschnitten aus dem Trainingsset, die mit erweiterten Bildern (Änderungen von Position, Drehung, Helligkeit, Detailgrad, ...) ergänzt wurden, um mehr Beispiele für die Bildübereinstimmung und damit positive Beispiele zu erhalten. Die Testdaten wurden nach einem ähnlichen Verfahren erstellt, allerdings aus dem ursprünglichen Testsatz.

Die Ergebnisse sowohl der Trainings- als auch der Testphase des Modells sind in der folgenden Tabelle dargestellt (es werden nur die besten Ergebnisse für eine bestimmte Modellstruktur gezeigt):

			Training		Testing	
Model	**Batch size**	**Epochs training**	**Loss**	**Accuracy**	**Loss**	**Accuracy**
strip_mse	20	600	0,0007	0,9994	0,0425	0,9481
strip_closs	20	600	0,0007	0,9994	0,0400	0,9484
mse	40	200	0,0041	0,9957	0,0561	0,9365
closs	20	400	0,0073	0,9933	0,0765	0,9104

Diese 4 Varianten von Modelleinstellungen wurden dann verwendet, um die Gesamtfähigkeit der gesamten Anwendung zur Erkennung des Standorts anhand des angegebenen Bildes zu untersuchen.

4.3 Erkennung von spezifischen Objekten

Zum Testen der Gesamtfähigkeiten der Anwendung wurden die Objektausschnitte in den Trainingsbildern durch den "Turm"-Teil des Siam-Netzes verarbeitet und der daraus resultierende Fingerabdruck in der Datenbank gespeichert. Wenn ein unbekanntes Bild vorgelegt wurde, wurden die Fingerabdrücke in der Datenbank vom Siamese-Netzwerk mit den vom YOLOv5-Netzwerk erhaltenen Bildausschnitten verglichen. Die unbekannten Bilder stammten aus dem ursprünglichen Testsatz.

So wurde für jeden unbekannten Ausschnitt der Ausschnitt aus derselben Kategorie mit dem höchsten Ähnlichkeitsgrad und einem Maß für diese Ähnlichkeit ermittelt. Die Menge dieser Ausschnitte aus dem Trainingssatz wurde dann abschließend verarbeitet, um den Standort des unbekannten Bildes zu bestimmen. Für diese Verarbeitung wurden vier verschiedene Methoden getestet:

Standortwahl nach maximaler Summe der Ähnlichkeitsmaße nach Standort (max_sum_qual)

- Standortauswahl nach dem ersten absolut höchsten Ähnlichkeitsmaß je Standort (max_abs_qual_first)

- Auswahl eines Ortes nach der maximalen Anzahl der Ähnlichkeiten je Ort (max_sum_cnt)

- Auswahl eines Standorts nach der maximalen Summe der Ähnlichkeitsmaße pro Standort unter Berücksichtigung der Gewichtung der Objektkategorien entsprechend ihrer Häufigkeit (max_sum_qual_wtd)

- Auswahl eines Standorts nach der maximalen Summe der Ähnlichkeitsmaße pro Standort unter Berücksichtigung der Gewichtung der Objektkategorien nach dem Kehrwert ihrer Häufigkeit (max_sum_qual_wtd_inv)

Die Ergebnisse der Gesamtqualität (Genauigkeit) der Standortbestimmung für die verschiedenen Modelleinstellungen und die verschiedenen endgültigen Methoden sind in der folgenden Tabelle dargestellt:

Model	**max_sum_qual**	**max_abs_qual_first**	**max_sum_cnt**	**max_sum_qual_wtd**	**max_sum_qual_wtd_inv**
mse	0,4084	0,2939	0,4008	0,3779	0,3588
closs	0,5038	0,4160	0,5191	0,4809	0,4351
strip-mse	0,6031	0,5458	0,5916	0,5611	0,4809
strip-closs	0,6298	0,5725	0,5916	0,6107	0,5153

Aus den erzielten Ergebnissen wird deutlich, dass es eine gewisse Diskrepanz zwischen den Ergebnissen der Tests am Siamesischen Netzmodell selbst und den Endergebnissen der gesamten Anwendung gibt. Der Grund dafür könnte in der ungeeigneten Zusammensetzung des Validierungssatzes während der Anwendungstests liegen. Die Verwendung des ursprünglichen Testsets könnte dazu geführt haben, dass auch Bilder von Orten getestet wurden, die im ursprünglichen Trainingsset nicht vertreten waren. Ein geeigneterer Ansatz wäre dann die Verwendung eines Satzes von Bildern von Standorten, die im Trainingssatz vertreten sind, aber völlig getrennt von diesem Satz aufgenommen wurden

Aus der Tabelle ist jedoch ersichtlich, dass die besten Ergebnisse mit dem Modell "strip_closs" und der endgültigen Klassifizierung nach der maximalen Summe der Ähnlichkeitsmaße nach Ort (max_sum_qual) erzielt wurden.

REFERENZEN

[1] L. Jose, Architectural Heritage Elements image Dataset, available online, https://old.datahub.io/dataset/architectural-heritage-elements-image-dataset, 2017.

[2] J. Llamas, P.M. Lerones, R. Medina, E. Zalama, J. Gómez-García-Bermejo, Classification of architectural heritage images using deep learning techniques Appl. Sci., 7 (10) (2017), p. 992, 10.3390/app7100992

View in ScopusGoogle ScholarL. Jose, Architectural Heritage Elements image Dataset, available online, https://old.datahub.io/dataset/architectural-heritage-elements-image-dataset, 2017.

[3] J. Fesl, J. Jelínek, K. Horníčková, Z. Nevařilová, M. Konopa, M. Feslová, AI-based system for cultural heritage objects identification from real photos, 12th International Conference on Advanced Computer Information Technologies (2022), pp. 476-479, 10.1109/ACIT54803.2022.9912752, 2022.

[4] H. Wölfflin Principles of art history, Eric Fernie, Art History and Its Methods. A Critical Anthology, Phaidon Press, New York - London (1915), pp. 137-140 (reprint) 2008.

[5] H. Focillon. The Life of Forms in Art (1934), pp. 171-173.

[6] E.H. Gombrich, StyleD. Preziosi (Ed.), The Art of Art History, A Critical Anthology, Oxford University Press, Oxford – New York (1998), pp. 150-163.

[7] G. Morelli, Italian painters Eric Fernie, Art History and Its Methods. A Critical Anthology, Phaidon Press, New York - London (1890), p. 108, (reprint) 2008.

[9] P. Chotěbor Stavební postupy svatovítské stavební huti ve středověku zjištěné při konzervaci katedrály sv. Víta na Pražském hradě, Staletá Praha, 22 (2016), pp. 122-129.

Metodika byla vytvořena v rámci projektu "Informační systém pro středověké památky v česko-bavorském příhraničí, č. 335, který je financován z Programu přeshraniční spolupráce Česká republika - Svobodný stát Bavorsko Cíl EÚS 2014 - 2020.