HTWK

Hochschule für Technik,
Wirtschaft und Kultur Leipzig

Bachelorarbeit
zur Erlangung des akademischen Grades

Bachelor of Science (B.Sc.)

im Bachelorstudiengang Medieninformatik
der Fakultät Informatik und Medien

Entwicklung eines Bilderkennungssystems zum finden von Duplikaten in einer Bilddatenbank bei der Registrierung neuer Bilder

vorgelegt von
Martino Thomann

Leipzig, den 11. September 2023

Erstprüfer:	Prof. Dr. Sibylle Schwarz
Zweitprüfer:	B.Sc. Marc Bellmann

Zusammenfassung

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Bei Spreadshirt werden täglich tausende Designs hochgeladen. Viele davon in der Partner-Area, wo Designer ihre Bilder zum Verkauf anbieten können. Dabei muss seitens Spreadshirt sichergestellt werden, dass die hochgeladenen Designs nicht verfassungsfeindlich sind und auch nicht gegen das Urheberrecht oder Spreadshirts eigene Richtlinien verstoßen. Die manuelle Überprüfung von so vielen Designs ist sehr aufwändig und häufig kommt es vor, dass die gleichen Designs mehrmals hochgeladen werden. Eine automatische Sperrung, von bereits verbotenen Designs, die erneut hochgeladen werden, kann den Überprüfungsprozess entlasten. Daher ist ein System notwendig, dass bei neu hochgeladenen Designs Duplikate in der Datenbank der verbotenen Designs erkennen kann.

1.2 Aufgabenstellung

Ziel dieser Bachelorarbeit ist die Entwicklung eines Systems, dass in der Lage ist, bei neuen Bildern zu erkennen, ob diese bereits in einer Datenbank mit bereits gespeicherten Bildern auftauchen. Ein neues Bild soll auch dann als Duplikat erkannt werden, wenn der Bildinhalt im Vergleich zum Original transformiert, also rotiert, skaliert, verschoben oder gespiegelt wurde. Fälle in denen das Bild anders gefärbt ist oder als Teil eines größeren Bildes auftaucht, sollen ebenfalls berücksichtigt werden.

Zur Implementierung des Systems soll ein "feature-based"(dt. merkmalbasierter) Algorithmus verwendet werden. Es gibt eine Vielzahl an merkmalbasierten Algorithmen, die jeweils ihre eigenen Stärken und Schwächen haben. Eine Auswahl dieser Algorithmen sollen innerhalb der Arbeit für den Anwendungsfall bei Spreadshirt getestet und miteinander verglichen werden.

1.3 Erfolgs- und Qualitätskriterien

Die Güte des Bilderkennungssystems soll anhand von Testdatensätzen ermittelt werden. Die Testdatensätze sind dabei in zwei Teilsätzen unterteilt. Der erste Teilsatz an Bildern stellt die Menge an gespeicherten Datenbankbildern dar. Der zweite Teilsatz enthält eine Untermenge an Duplikaten aus dem Datenbanksatz und eine Menge an neuen Bildern, die nicht im Datenbanksatz auftauchen. Getestet wird in verschiedenen Szenarien, die die Robustheit des Systems gegenüber bestimmter Sonderfälle testen soll. Je nach Szenario sind die Duplikate auf unterschiedliche Weise im Verglich zum Original verändert. Als Vergleich dient der pHash-Algorithmus, der momentan bei Spreadshirt zur Duplikatensuche verwendet wird.

Die verwendeten Metriken werden in den Grundlagen 2.2 erklärt. Am wichtigsten ist dabei ein hoher Recall, sodass möglichst viele Duplikate durch das System abgefangen werden. Da automatisch gesperrte Designs nochmal manuell geprüft werden, fällt eine niedrigere Spezifizität bei der Auswertung nicht so sehr ins Gewicht. Laufzeit- und Speicherkosten sollen ebenfalls innerhalb der Arbeit abgeschätzt werden.

Um für die Verwendung bei Spreadshirt in Frage zu kommen, muss das neue System zuverlässiger sein, als die momentan eingesetzte pHash-Implementaion. Dafür wird ein höherer Recall angestrebt. Um sicherzustellen, dass Spezifizität nicht zu sehr absinkt, wird auf eine höhere oder zumindest gleichbleibende Balancierte-Genauigkeit abgezielt.

2 Grundlagen

2.1 Scale Invariant Feature Transform

2.2 Test-Metriken

Das Bilderkennungssystem soll die Bilder in zwei Klassen einordnet: Duplikate und nicht-Duplikate/neue Bilder. Da es in diesem Fall bei der Klassifizierung nur zwei mögliche Klassen gibt, kann man das Bilderkennungssystem als binären Klassifikator bezeichnen.

Die Performance von binären Klassifikatoren kann durch die Werte einer Wahrheitsmatrix quantifiziert werden. Die Wahrheitsmatrix gibt dabei an, wie viele richtige und falsche Entscheidungen das System bei der Klassifikation getroffen hat. [Tha20, S. 170]

	Wahre Klassifikation	
System Klassifikation:	Duplikat	nicht-Duplikat
Duplikat	Anzahl wahre Duplikate (TP)	Anzahl falsche Duplikate (FP)
nicht-Duplikat	Anzahl falsche nicht-Duplikate (FN)	Anzahl wahre nicht-Duplikate (TN)

Aus den Werten der Wahrheitsmatrix lassen sich weitere Metriken ableiten, die für die Bewertung eines binären Klassifikators nützlich sein können. Interessant für diese Arbeit sind Recall, Spezifizität und Balancierte-Genauigkeit.

Der Recall, oder auch true positive rate, gibt das Verhältnis zwischen den Duplikaten, die das System korrekt klassifiziert hat, und allen Duplikaten, die sich in dem Suchdatensatz befinden, an. [Tha20, S. 172]

$$Recall = \frac{TP}{TP + FN}$$

Die Spezifizität, oder auch true negative rate, gibt das Verhältnis zwischen den nicht-Duplikaten, die das System korrekt klassifiziert hat, und allen nicht-Duplikaten, die sich im Suchdatensatz befinden an. [Tha20, S. 172]

Die Accuracy (dt. Genauigkeit), gibt das Verhältnis zwischen den Bildern, die das System korrekt klassifiziert hat, und allen Bildern im Suchdatensatz an. Die Genauigkeit spiegelt die Fähigkeit des Systems Bilder richtig zu Klassifizieren wieder.[Tha20, S. 171]

Allerdings ist die Genauigkeit kein zuverlässiger Wert, wenn man mit unausgeglichenen Datensätzen arbeitet. Ein Datensatz gilt dann als unausgeglichen, wenn einer Klassifikation mehr Elemente angehören, als der anderen Klassifikation. [Tha20, S. 171] In den Testdatensätzen, die für diese Arbeit verwendet werden, befinden sich mehr nicht-Duplikate als Duplikate. Dieses Ungleichgewicht kann die Genauigkeit stark beeinflussen. So kann zum Beispiel eine gute Spezifizität einen schlechten Recall ausgleichen, wenn der Datensatz zum größten Teil aus nicht-Duplikaten besteht.

Daher kommt bei unbalancierten Datensätzen die balanced-accuracy (dt. Balancierte-Genauigkeit) zum Einsatz. [Tha20, S. 175] Da sie den Durchschnitt aus Recall und Spezifizität bildet, werden Unterschiede zwischen den beiden Werten ausgeglichen. Dadurch ist die Balancierte-Genauigkeit robust gegenüber unausgeglichenen Datensätzen.

Literaturverzeichnis

[Tha20] Alaa Tharwat. Classification assessment methods. *Applied computing and informatics*, 17(1):168–192, 2020.

Abbildungsverzeichnis

Tabellenverzeichnis

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Graduierungsarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe. Alle den benutzten Quellen wörtlich oder sinngemäß entnommene Stellen sind als solche einzeln kenntlich gemacht.

Diese Arbeit ist bislang keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht worden.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Leipzig, 11. September 2023 Unterschrift