



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 |
| 1.1 | Motivation | 1 |
| 1.2 | Aufgabenstellung | 1 |
| 1.3 | Erfolgs- und Qualitätskriterien | 2 |
| | Literaturverzeichnis | I |
| | Abbildungsverzeichnis | II |
| | Tabellenverzeichnis | III |

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. merkmalsbasierter) Algorithmus verwendet werden. Es gibt eine Vielzahl an merkmalsbasierten 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 Vergleich zum Original verändert. Als Vergleich dient der pHash-Algorithmus, der momentan bei Spreadshirt zur Duplikatensuche verwendet wird.

Die verwendeten Metriken werden in den Grundlagen 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 Spezifität 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 Spezifität nicht zu sehr absinkt, wird auf eine höhere oder zumindest gleichbleibende Balancierte-Genauigkeit abgezielt.

Literaturverzeichnis

- [Low99] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.
- [Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

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