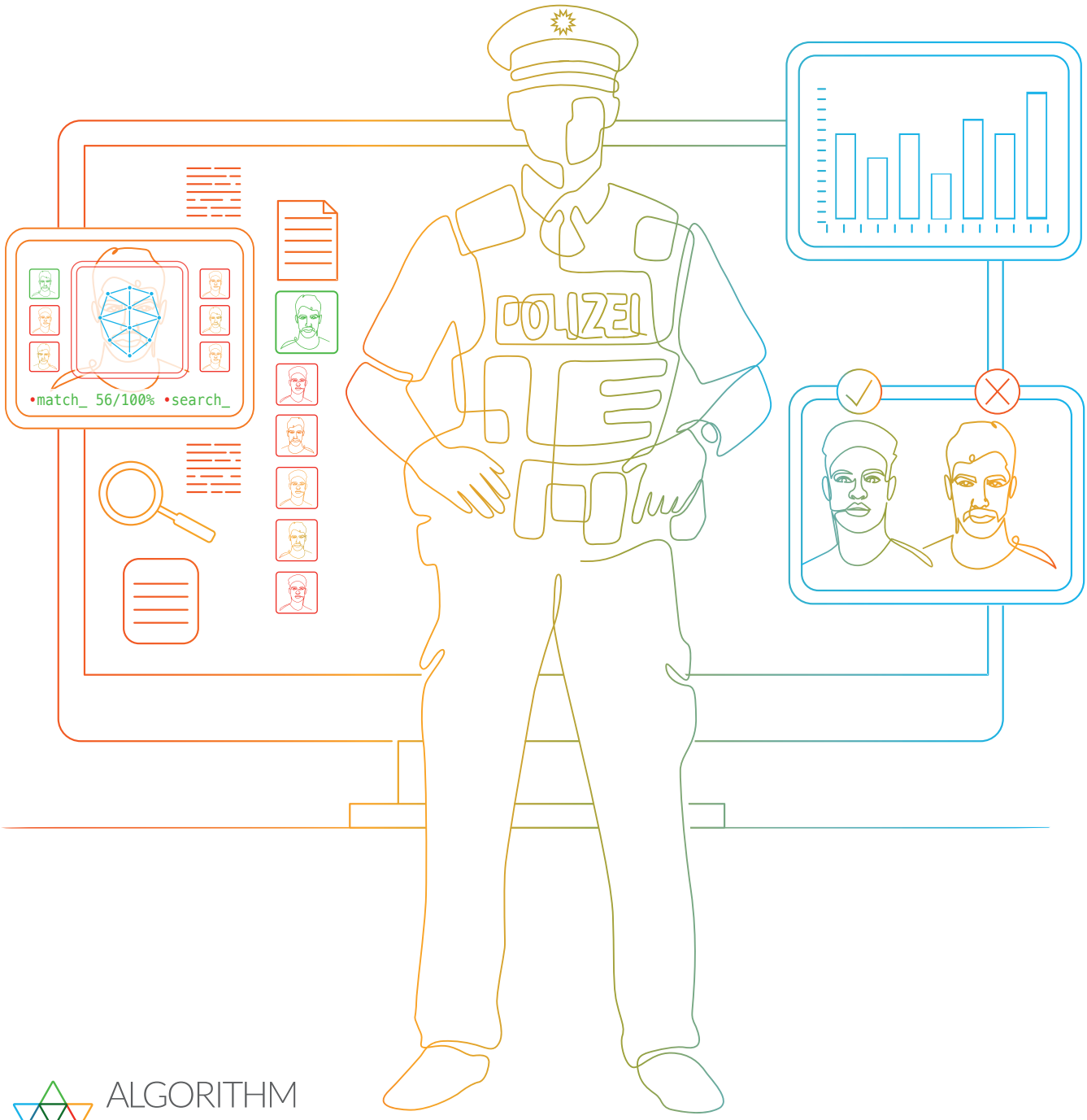
Braucht die Polizei eine Datenbank zum biometrischen Abgleich? Das Durchsuchen von Internetbildern zur Gesichtserkennung

Technisches Gutachten
von Prof. Dr. Dirk Lewandowski

September 2025



ALGORITHM WATCH

/ Zusammenfassung

Ist ein biometrischer Abgleich zwischen Aufnahmen von Überwachungskameras und im Internet verfügbaren Bildern ohne Verwendung einer Datenbank möglich?

Die KI-Verordnung der EU verbietet es ausnahmslos, durch ein anlassloses Scraping von Gesichter-Aufnahmen Datenbanken zur Gesichtserkennung aufzubauen. Insofern würden nationale Gesetzesvorhaben, die einen biometrischen Abgleich mit Bildern aus dem Internet vorsehen, geltendem EU-Recht zuwiderlaufen, falls dieser Abgleich nur mithilfe von Datenbanken stattfinden kann.

Bilder-Suchmaschinen beantworten Anfragen (wie Suchmaschinen im Allgemeinen) nicht „live“, sonst könnten die Anfragen erst nach extrem langer Wartezeit beantwortet werden. Die Suchmaschinen durchsuchen einen vorhandenen Datenbestand, der zuvor aus dem Internet gesammelt, für eine anschließende Suche aufbereitet und in einer Datenbank gespeichert wurde. Bilder müssen vorverarbeitet werden, um nach ihnen für einen Abgleich suchen zu können. Die Vorverarbeitung ist so komplex, dass sie nicht erst durchgeführt werden kann, wenn eine Suchanfrage gestellt wird. Erst das Sammeln und Vorverarbeiten der Daten ermöglicht eine effiziente Suche. Für den Zugriff auf die vorverarbeiteten Bilder ist es notwendig, sie in einer Datenbank zu speichern.

Es ist also technisch nicht umsetzbar, frei verfügbare Bilder aus dem Internet für einen Abgleich praktikabel durchsuchbar zu machen, ohne eine Datenbank zu erstellen, wie das Gutachten zeigt, das AlgorithmWatch bei Prof. Dr. Dirk Lewandowski von der Hochschule für Angewandte Wissenschaften Hamburg in Auftrag gegeben hat.

/ Inhalt

1 Einleitung ... 4

2 Wie lassen sich große Datenmengen im Internet sinnvoll erfassen? ... 5

2.1 Kann das Web „live“ durchsucht werden? ... 5

2.2 Arbeitsweise von Suchmaschinen ... 6

2.2.1 Auffinden und Erfassen der Inhalte (Crawling) ... 7

2.2.2 Vorverarbeitung der Inhalte (Indexierung) ... 8

2.2.3 Abgleich von Suchanfragen mit dem Datenbestand (Ranking) ... 9

2.3 Arbeitsweise von Bildersuchmaschinen ... 10

2.3.1 Allgemeine Bildersuchmaschinen ... 10

2.3.2 Biometrische Bildersuchmaschinen: Beispiel PimEyes ... 12

2.3.3 Potenziell für den Abgleich geeignete, offene Bilderbestände ... 13

3 Datenbanken ... 14

4 Abgleich biometrischer Daten ohne Verwendung einer Datenbank ... 15

4.1 Auf der Seite der anfragenden Stelle vorhandene Daten ... 15

4.2 Live-Suche im Web ... 16

4.2.1 Web-Crawling mit Einzelabgleich aller Bilder ... 16

4.2.2 Focused Crawling ... 17

4.3 Zugriff über Application Programming Interfaces (APIs) ... 17

4.4 Scraping von Bildersuchmaschinen ... 17

5 Fazit ... 18

6 Literaturverzeichnis ... 19

7 Kurzbiografie des Autors ... 20

1 Einleitung

Dieses Gutachten wurde von der AW AlgorithmWatch gGmbH in Auftrag gegeben und vom Verfasser ohne über die im Folgenden benannte Fragestellung hinausgehende Weisungen erstellt.

Das Gutachten setzt sich mit der Frage auseinander, ob und ggf. inwieweit es technisch möglich ist, biometrische Daten zu Gesichtern, die Behörden vorliegen, mit öffentlich zugänglichen Daten aus dem Internet mittels einer automatisierten Anwendung zur Datenverarbeitung biometrisch abzugleichen, ohne eine Datenbank zu erstellen. Hierbei ist davon auszugehen, dass die Stelle, die den Abgleich vornehmen möchte, nur anhand des biometrischen Gesichtsmusters einer Person sucht, d.h. keine weiteren biometrischen Daten wie beispielsweise Stimmproben vorliegen.

Das Gutachten gliedert sich in drei Hauptteile: In Abschnitt 2 wird grundlegend dargestellt, welche Datenmengen im Web vorhanden sind, welche Ansätze zu ihrer Erfassung verfolgt werden und wie insbesondere Suchmaschinen und spezialisierte Bildersuchmaschinen funktionieren. Hierbei wird gezeigt, dass sowohl in der textuellen als auch in der Bildersuche eine Vorverarbeitung nötig ist, um einen späteren Abgleich von Suchanfragen und Informationsobjekten zu ermöglichen.

In Abschnitt 3 wird beschrieben, was eine Datenbank konstituiert und wie die Datenbasis einer Datenbank entsteht. Hierbei wird gezeigt, dass eine Datenbank entsteht, sobald mehr als ein Informationsobjekt systematisch gespeichert wird, und dass die Speicherdauer für das Zustandekommen einer Datenbank unerheblich ist.

Abschnitt 4 beschäftigt sich mit den theoretischen Möglichkeiten des Abgleichs eines vorliegenden Bilds inklusive seines biometrischen Templates mit frei verfügbaren Web-Daten. Dabei wird gezeigt, dass es mit keinem Ansatz möglich ist, ohne die Erstellung einer Datenbank Bilder aus dem Web praktikabel in einer Form durchsuchbar zu machen, die den gewünschten Abgleich ermöglicht.

2 Wie lassen sich große Datenmengen im Internet sinnvoll erfassen?

Wenn der offene Datenbestand des Web durchsucht werden soll, ist zunächst einmal auf die grundlegenden Charakteristika dieses Datenbestands einzugehen:

1. **Die Daten liegen verteilt vor und es gibt kein zentrales Verzeichnis.**
2. **Die Daten verändern sich ständig: Informationsobjekte werden hinzugefügt, geändert und gelöscht.**
3. **Die Größe des Datenbestands ist enorm.**

Die Informationsobjekte im Web werden als Dokumente bezeichnet. Bei einem Dokument kann es sich unter anderem um einen Text, ein Bild oder ein Video handeln. Der Begriff Dokument beschränkt sich also nicht auf einen bestimmten Typ von Informationsobjekt.

Dokumente im Web können von jeder Person erstellt und veröffentlicht werden, ohne dass eine Form der Registrierung in einem zentralen Verzeichnis nötig ist. Die Daten liegen außerdem verteilt auf verschiedenen Servern. Das bedeutet, dass jedes Suchsystem erst einmal selbständig die im Web vorhandenen Daten auffinden muss, unabhängig davon, ob das Web vollständig (Universalsuchmaschinen) oder nur in Teilen (Spezialsuchmaschinen) erfasst werden soll.

Die Dynamik des Web bedeutet, dass Suchsysteme die Inhalte des Web beständig auf Veränderungen überprüfen müssen. Der Datenbestand einer Suchmaschine ist dabei ein möglichst vollständiges und aktuelles Abbild der relevanten Inhalte des Web (Lewandowski, 2021, S. 33). Universalsuchmaschinen wie Google machen nur diesen aktuellen Datenbestand durchsuchbar, d.h. dass bspw. ein Bild, das aus dem Web entfernt wurde, auch aus dem Datenbestand der Suchmaschine entfernt wird, sobald dies von der Suchmaschine festgestellt wird. Allerdings gibt es auch Suchmaschinen, die gelöschte Informationsobjekte weiterhin in ihrem Datenbestand behalten (Archivsuchmaschinen, z.B. *Wayback Machine*).

Die Größe des Web stellt Anbieter von Suchsystemen vor enorme Herausforderungen. Aufgrund des Fehlens eines zentralen Registers aller Webseiten und der Dynamik des Web ist eine Messung der Größe des Web schwierig, allerdings sind gerade von Google als weltweit größter Suchmaschine zumindest ältere Daten bekannt: Schon im Jahr 2016 waren Google 130 Billionen URLs bekannt (Schwartz, 2016). Dazu gehören auch nicht mehr aktive URLs und Seiten, die nicht in Googles Datenbestand aufgenommen wurden. Googles durchsuchbarer Datenbestand lag im Jahr 2020 bei 400 Milliarden Dokumenten (Shepard, 2023). Googles Spezialsuchmaschine für Bilder hatte im Jahr 2010 bereits einen Datenbestand von mehr als zehn Milliarden Bildern (Smith et al., 2010) und dürfte seither erheblich gewachsen sein.

2.1 Kann das Web „live“ durchsucht werden?

Ein häufiges Missverständnis über den Aufbau von Suchmaschinen und die Suche im Web ist, dass eine Suchmaschine in dem Moment, in dem von einer Person eine Suchanfrage eingegeben wird, das Web durchsucht, also eine Art „Live-Suche“ vornimmt. Dies ist nicht der Fall. Vielmehr durchsucht eine Suchmaschine ihren eigenen Datenbestand, der zuvor aus dem Web gesammelt, für die Suche aufbereitet und in einer Datenbank gespeichert wurde. Der Grund liegt darin, dass eine „Live-Suche“ im Web nicht möglich ist: Eine Suchmaschine müsste (1) über die Verlinkungen zwischen Dokumenten jedes Dokument im Web auffinden und (2) jedes dieser Dokumente mit der Suchanfrage abgleichen. Abgesehen von der Ineffizienz dieses Vorgehens und des riesigen Datenverkehrs, der damit einhergehen würde, ließen sich Suchanfragen so nicht in einer

vertretbaren Zeit beantworten. Bedenkt man, dass das Web aus hunderten Milliarden Dokumenten besteht, wird klar, dass sich Suchanfragen so nicht beantworten lassen. Eine informationssuchende Person müsste selbst bei der Verwendung moderner Rechner jahrelang auf eine Antwort auf ihre Suchanfrage warten. Dazu kommt, dass sich der Datenbestand im Web beständig ändert, was zur Folge hat, dass sich während eines solchen hypothetischen Suchdurchlaufs zwischenzeitlich der Datenbestand erheblich verändert hätte, was selbst im günstigsten Fall zu einer unvollständigen und veralteten Antwort führen würde. Damit müsste die Suche unmittelbar wieder von vorn beginnen. Es ist also offensichtlich, dass eine solche „Live-Suche" durch die Inhalte des Web kein gangbarer Weg ist.

Das beschriebene Vorgehen der „Live-Suche" lässt sich mit der Suche nach einem Stichwort in Büchern vergleichen. Wenn man in der eigenen Bibliothek alle Nennungen eines Stichworts finden möchte, müsste man bei einem vergleichbaren Vorgehen jedes Buch von vorne bis hinten durchlesen und jedes Auffinden des Stichworts notieren. Man müsste also alle Bücher *vollständig* lesen, weil sonst unklar bliebe, ob es noch ein weiteres Vorkommen des gesuchten Stichworts gibt. Die Suche ist also erst vollständig, wenn alle Bücher vollständig gelesen wurden. Es ist offensichtlich, dass ein solches Vorgehen nicht zielführend und schon gar nicht effizient ist. Gelöst wird das Problem durch Register, die von Stichwörtern auf Seitenzahlen verweisen. Ähnlich arbeiten auch Suchmaschinen und andere Suchsysteme: Sie erstellen Indexe (Register), also geordnete Wortlisten, die auf die Dokumente verweisen, in denen das jeweilige Wort vorkommt. Es wird also eine Datenbank mit den Stichwörtern und den zugehörigen Fundstellen erstellt. Nur so kann eine effiziente Verarbeitung von Suchanfragen erreicht werden.

Um ein praktikables Suchsystem zu erstellen, bedarf es also einer umfangreichen Vorverarbeitung der Daten. Diese besteht im Fall der Suche im Web aus der Sammlung der Daten aus externen Quellen (Abschnitt 2.2.1) sowie der Aufbereitung dieser Daten (Abschnitt 2.2.2).

2.2 Arbeitsweise von Suchmaschinen

Im Folgenden werden der Aufbau und die Arbeitsweise von Suchmaschinen, aufbauend auf der detaillierten Darstellung in Lewandowski (2021), beschrieben. Der Schwerpunkt wird dabei auf das Erfassen der Inhalte (Crawling), die Vorverarbeitung der gefundenen Inhalte (Erstellung von Datenbanken zur Ablage der Dokumente und der Indexe) und den Abgleich zwischen Suchanfragen und Datenbestand gelegt.

Eine Suchmaschine wird folgendermaßen definiert: „Eine Suchmaschine (auch: Web-Suchmaschine; Universalsuchmaschine) ist ein Computersystem, das verteilte Inhalte aus dem World Wide Web mittels Crawling erfasst und über eine Benutzerschnittstelle durchsuchbar macht, wobei die Ergebnisse in einer nach systemseitig angenommener Relevanz geordneten Darstellung aufgeführt werden." (Lewandowski, 2021, S. 29)

Grundlegend besteht eine Suchmaschine aus zwei Prozessen: dem *indexing process* und dem *query process* (Croft et al., 2009). Im *indexing process* werden die Informationsobjekte aufgefunden und so vorverarbeitet, dass sie effizient durchsucht werden können. Im query process werden Suchanfragen verarbeitet und passende Informationsobjekte ausgegeben. Eine wichtige Unterscheidung ist, dass im *indexing process* eine umfassende Vorverarbeitung stattfindet, die in den meisten Fällen nicht zeitkritisch ist. Im Gegensatz dazu müssen im *query process* Suchanfragen schnell verarbeitet werden, da Nutzende die Ergebnisse in einer vertretbaren Zeit erwarten. Die Erwartungen können hier je nach System sehr unterschiedlich sein: In der Web-Suche werden Ergebnisse in der Regel in weniger als einer Sekunde erwartet, bei komplexeren Analysen von Web-Daten, beispielsweise in agentenbasierten KI-Systemen (u.a. „Deep Research" in ChatGPT) werden auch Wartezeiten von einigen Minuten akzeptiert. In der professionellen Recherche können durchaus auch weit längere Wartezeiten akzeptabel sein; eine detailliertere Verarbeitung kann auch zu einer erheblichen Steigerung der Ergebnisqualität führen (Teevan et al., 2014).

Durch die Vorverarbeitung im *indexing process* wird erreicht, dass im *query process* nicht mehr die Inhalte des Web, sondern die von der Suchmaschine in Form einer Index-Datenbank aufbereiteten Web-Daten durchsucht werden. Gleiches gilt für alle anderen Systeme, die eine Suche in ihren Datenbeständen erlauben: Um eine funktionsfähige Suchfunktion anbieten zu können, werden die Daten vorverarbeitet. Wird eine Suchanfrage gestellt, werden die Repräsentationen der Informationsobjekte durchsucht, nicht die Informationsobjekte selbst. Wenn man also beispielsweise die Suchfunktion in einem sozialen Netzwerk wie LinkedIn nutzt, um nach einer Person zu suchen, wird der eingegebene Name mit einem Index (Register) der in der Datenbank vorhandenen Namen abgeglichen und entsprechend werden die Namen, die der Suchanfrage entsprechen, ausgegeben.

2.2.1 Auffinden und Erfassen der Inhalte (Crawling)

Das *Web Crawling* bildet das Rückgrat jeder Suchmaschine und ist verantwortlich für die systematische Erfassung von Informationen aus dem Web. Das verteilte Crawlingsystem agiert kontinuierlich, um den dynamischen Charakter und die immense Größe des Web abzubilden. Der Hauptzweck des Crawlings besteht darin, Webdokumente zu identifizieren und in den Datenbestand der Suchmaschine herunterzuladen.

Der Web-Crawler ist eine Softwarekomponente, die autonom das Web durchsucht. Der Prozess beginnt mit einer Initialmenge von URLs, dem sogenannten *seed set*. Aus diesen initialen Dokumenten werden alle Links auf andere Dokumente extrahiert. Im weiteren Verlauf werden diese URLs besucht, die dort hinterlegten Dokumente heruntergeladen und aus diesen wiederum die Links extrahiert. In der Theorie können mit dieser Methode alle im Web vorhandenen Dokumente gefunden werden, sofern sie verlinkt sind. Der Crawler baut also in dem Prozess des Besuchs immer weiterer URLs ein Netz von Dokumenten auf, das im Idealfall der Struktur des aktuell bestehenden Web entspricht. Damit erstellt der Crawler eine *Kopie* des realen Web.

Zentrale Komponenten eines Web-Crawlers sind die Warteschlange *(frontier)*, der Downloader, der Parser und der Content Store.

- In der Warteschlange wird eine Liste der noch zu erfassenden URLs geführt. Jede URL, die in einem Dokument gefunden wurde, wird der Warteschlange hinzugefügt. Die Warteschlange kann nach unterschiedlichen Kriterien (bspw. Popularität der Websites) priorisiert werden.

- Der Downloader lädt das gefundene Dokument herunter, sofern es für die Suchmaschine neu ist oder seit dem letzten Besuch des Crawlers verändert wurde.

- Der Parser extrahiert alle Links aus den gefundenen Dokumenten und fügt sie der Warteschlange hinzu. Da auch Bilder in die HTML-Dokumente mittels Links eingebettet werden, werden auch diese Verweise entsprechend erfasst. Sie können auch der Warteschlange eines spezialisierten Bilder-Crawlers hinzugefügt werden (s. Abschnitt 2.3.1).

- Im Content Store werden schließlich die heruntergeladenen Dokumente gespeichert. Der Content Store ist der zentrale Speicher aller Dokumente, die allerdings, bevor sie durchsuchbar sind, umfangreich vorverarbeitet werden müssen.

Für jede URL muss also folgender Prozess durchgeführt werden: (1) Eine URL wird aus der Warteschlange ausgewählt. (2) Der Downloader lädt das entsprechende Dokument herunter und speichert es in einer Datenbank, dem Content Store. (3) Der Parser extrahiert alle Links und weitere relevante Informationen aus dem Dokument. (4) Die neu entdeckten Links werden hinsichtlich ihrer Qualität und Priorität bewertet. (5) Die qualifizierten und priorisierten Links werden der Warteschlange hinzugefügt.

Dieser Zyklus wird ständig wiederholt, wodurch der Crawler das Web „durchwandert" und neue sowie aktualisierte Inhalte entdeckt, die der Content-Store-Datenbank hinzugefügt werden. Aufgrund der Größe des Web und seiner beständigen Aktualisierung sind die Web-Crawler der großen Suchmaschinen als verteilte Systeme mit tausenden von parallel arbeitenden Rechnern konzipiert. Ansonsten wäre die Bewältigung der enormen Datenmengen und der hohen Dynamik des Web nicht möglich.

Eine Konsequenz der Menge der Daten im Web und der enormen finanziellen und technischen Ressourcen, die für ihre beständige Erfassung benötigt werden, ist auch das Vorhandensein von nur vier umfassenden Web-Indexen, also Datenbanken des Web: Neben Google verfügen nur Microsoft (Bing), Yandex und Baidu über Datenbestände von einer Größe, die eine weitgehend vollständige Suche über die Web-Inhalte erlauben (Lewandowski, 2021, S. 194–197).

Es ist wichtig, dass jeder Crawler die sog. Crawling-Etikette befolgt. Wenn ein Crawler zu viele Anfragen in kurzer Abfolge an eine Website sendet, um bspw. alle Inhalte dieser Website herunterzuladen, kann das zu einer Überlastung der Server dieser Website kommen. Dadurch kann es zu einem Ausfall der Website kommen oder zumindest dazu, dass echte Nutzende die Dokumente der Website nur verzögert abrufen können. Während Website-Betreiber von dem Traffic, den sie durch Suchmaschinen erhalten, enorm profitieren und teils auch ihre Geschäftsmodelle darauf aufbauen, werden die Zugriffe durch andere Crawler oft als lästig betrachtet, da sie Ressourcen der Server verbrauchen, ohne eine Gegenleistung zu erbringen.

Website-Betreiber haben die Möglichkeit, Crawler zu steuern und auszuschließen. In der Datei robots.txt können Anweisungen für Crawler hinterlegt werden, beispielsweise auch zur Crawl-Frequenz (d.h. in welchem Abstand der Crawler Anfragen an den Server senden darf). Weiterhin können bestimmte Crawler vollständig ausgeschlossen werden. Von dieser Möglichkeit machen Website-Betreiber ausführlich Gebrauch, insbesondere, wenn sich Crawler in der Vergangenheit nicht an die Crawling-Etikette gehalten haben oder wiederholt viel Traffic verursachen. Aktuell ist eine enorme Zunahme des durch Crawling verursachten Traffics zu beobachten; vor allem, weil mittels Crawling auch Daten, die als Datenbasis für das Training von generativen Sprachmodellen dienen, gesammelt werden. Zusammenfassend lässt sich also sagen, dass das Crawling effizient geschehen muss und nicht die Ressourcen der von den Website-Betreibern betriebenen Server überlasten darf. Dieser Punkt ist von Bedeutung bei der Frage des Web-Crawlings mit Einzelabgleich aller Bilder (Abschnitt 4.2.1).

2.2.2 Vorverarbeitung der Inhalte (Indexierung)

Das Ergebnis des Crawlings ist eine Datenbank, die die Kopien aller gefundenen Dokumente enthält (Content Store). Um die Dokumente durchsuchbar zu machen, ist eine umfangreiche Vorverarbeitung nötig. Zuerst werden Dokumente vereinheitlicht und aufbereitet. Dazu gehören unter anderem die Konvertierung unterschiedlicher Formate in ein einziges Format (bspw. XML), die Konvertierung von unterschiedlichen Textkodierungen, das Erkennen von Dubletten, das Entfernen von nicht inhaltstragenden Elementen aus Dokumenten (bspw. Navigationselemente, Werbung), das Zerlegen von Text in Tokens (Wörter oder Wortbestandteile) und deren Vereinheitlichung, Vereinheitlichung von morphologischen Varianten von Wörtern (Stemming), das Erkennen von Mehrwortausdrücken und das Extrahieren von HTML-Markup (Croft et al., 2009). Die Vorverarbeitung der Dokumente ist also komplex und kann deshalb nicht erst durchgeführt werden, wenn eine Suchanfragen gestellt wird. Ebenso wie das Crawling muss die Vorverarbeitung auf einer Vielzahl von Rechnern parallel erfolgen, um die Datenmassen des Web verarbeiten zu können.

Vorverarbeitungen werden bei allen Datenbeständen eingesetzt. Um beispielsweise eine Bilderdatenbank mit biometrischen Informationen erstellen zu können, müssen in der Vorverarbeitung die biometrischen Merkmale aus den Bildern gewonnen werden.

Sie werden als Templates gespeichert. Templates können für Merkmale von Gesichtern erstellt werden, aber auch für andere biometrische Merkmale (bspw. Stimme, Gang, Iris, Fingerabdruck).

Auch die Vorverarbeitung führt noch nicht zu einem durchsuchbaren Datenbestand. Erst in der Indexierung werden Strukturen gebildet, die eine effiziente Suche erlauben. Ein Index entspricht einem Register und invertiert die Dokumente, die in der Content-Store-Datenbank gespeichert sind. Jedes (vorverarbeitete) Wort und die in der Vorverarbeitung erkannten Mehrwortausdrücke werden aus einem Dokument extrahiert und einer alphabetischen Liste mit ihrer Fundstelle hinzugefügt. In dieser Wortliste wird verzeichnet, in welchen Dokumenten das entsprechende Wort vorkommt. Diese Struktur sorgt dafür, dass bei der Suche nicht mehr die Dokumente selbst durchsucht werden, sondern die Suchbegriffe nur noch in der Wortliste nachgeschlagen werden müssen. Die Verweise auf die Dokumente *(pointer)* führen dann entweder zu den Dokumenten selbst – wenn diese im System selbst gespeichert werden – oder zu den URLs der Dokumente. Auch die Indexe müssen beständig aktualisiert werden, was insbesondere bei großen Datenmengen eine Herausforderung darstellt. Auch die Indexe liegen verteilt über eine Vielzahl von Rechnern vor, damit eine effiziente Suche möglich wird.

Die Indexierung ist ein Kernelement jeglicher Information-Retrieval-Systeme, unabhängig von ihrem konkreten Zweck. So lässt sich das für die Web-Suche Beschriebene auf Datenbanken mit biometrischen Informationen übertragen: Im Index finden sich Metadaten (bspw. Aufnahmedatum, Quelle/URL) und die biometrischen Repräsentationen (Templates) der Bilder. Die Content-Store-Datenbank enthält die Bilder selbst.

2.2.3 Abgleich von Suchanfragen mit dem Datenbestand (Ranking)

Das – vermeintlich – zentrale Element einer Suchmaschine ist der Abgleich der Suchanfragen mit dem Datenbestand und das damit verbundene Ranking der gefundenen Dokumente. Im Zusammenhang mit der umfassenden Leistung des Auffindens, der Erfassung und Vorverarbeitung der Dokumente wird allerdings deutlich, dass der Abgleich zwischen Suchanfragen und Dokumenten nur eine der Kernfunktionen einer Suchmaschine ist.

Suchmaschinen gleichen Suchanfragen und Dokumente mittels der erstellten Indexe ab, wobei jeweils die *Repräsentation der Suchanfrage* und die *Repräsentation der Dokumente* verglichen werden. Dabei werden die Suchanfragen analog der Vorverarbeitung der Dokumente vorverarbeitet, d.h. es wird auch von der Suchanfrage eine Repräsentation erstellt. Erst so wird ein sinnvoller Abgleich möglich. So muss bspw. die Erkennung von Mehrwortausdrücken auch für die Suchanfragen durchgeführt werden, da anderenfalls der Mehrwert, der durch die Erkennung dieser Ausdrücke in der Vorverarbeitung generiert wurde, in der Suche nicht ausgenutzt werden könnte. Da auch nicht textuelle Inhalte wie Bilder in der Form von Text repräsentiert werden, gilt die Vorverarbeitung der Suchanfragen hier analog.

Ein einfaches Ranking kann auf der Basis der statistischen Auswertung der Texte von Dokumenten erfolgen. Hier werden beispielsweise die Häufigkeit und die Platzierung der Suchbegriffe im Dokument (bspw. in Überschriften) gemessen. Allerdings stellt sich gerade bei Web-Inhalten, deren Qualität von hochwertigen Dokumenten bis hin zu Spam-Dokumenten reicht, die Herausforderung, neben der textlichen Passung auch die Qualität der Dokumente zu bewerten. Dies erfordert ein komplexes Set von weiteren Rankingfaktoren, die sich in die folgenden Gruppen unterteilen lassen (Lewandowski, 2021, S. 95ff.): (1) Popularität, (2) Aktualität, (3) Lokalität, (4) Personalisierung, (5) technische Rankingfaktoren.

Das Ergebnis einer Suchanfrage ist das Ranking der Ergebnisse, d.h. die gefundenen Ergebnisse werden in eine Reihenfolge gebracht. Die Annahme hinter jedem Ranking ist, dass die Dokumente nach absteigender Relevanz gereiht werden, d.h. der informationssuchenden Person werden die vom Suchsystem als am relevantesten angesehenen Dokumente zuerst angezeigt. Damit ist impliziert,

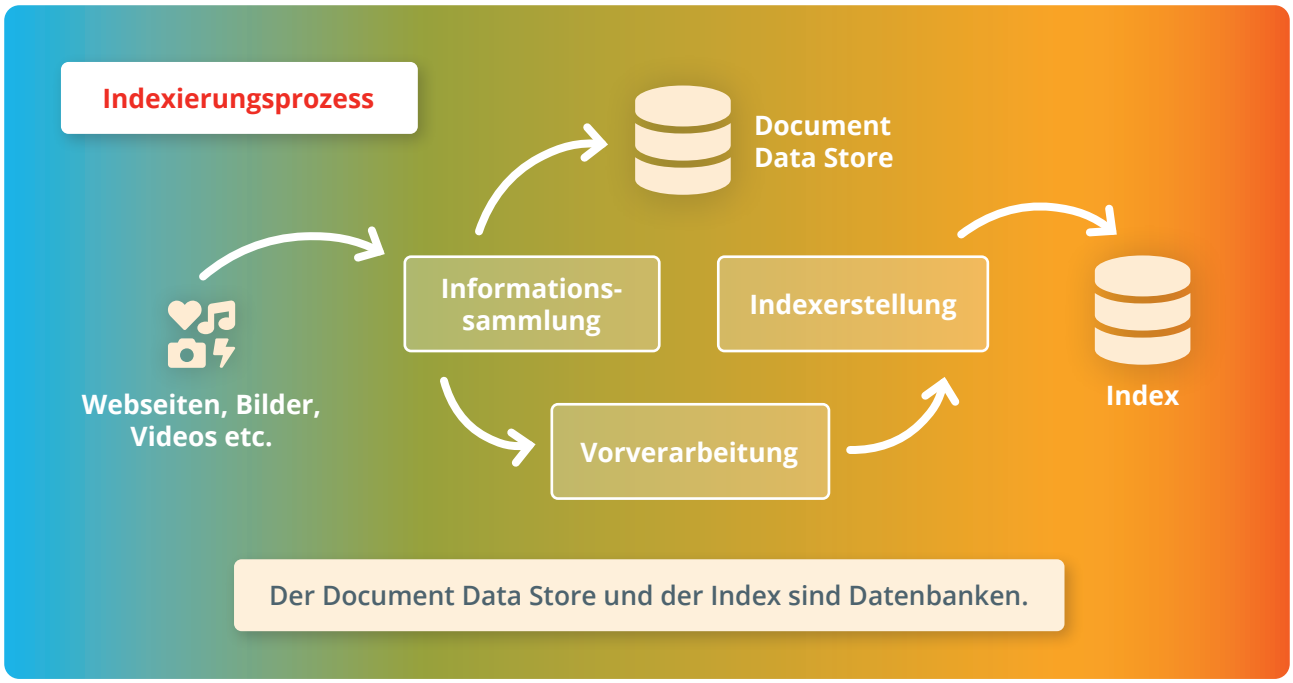


Abbildung 1: Indexierungsprozess nach Croft, Metzler & Strohman, 2009, https://ciir.cs.umass.edu/downloads/SEIRiP.pdf

dass es für die Beantwortung von Suchanfragen in der Treffermenge mehr und weniger gut geeignete Dokumente gibt, jedoch nicht absolut richtige oder falsche. Dies unterscheidet Rankings von exakten Abfragen in Datenbanken (s. Abschnitt 3), die eine klar bestimmte Treffermenge nach exakten Kriterien ergeben (bspw. die Zahl aller in Hamburg gemeldeten Personen unter 30 Jahren).

2.3 Arbeitsweise von Bildersuchmaschinen

2.3.1 Allgemeine Bildersuchmaschinen

Im vorangegangenen Abschnitt wurde die allgemeine Funktionsweise von Suchmaschinen beschrieben, um die grundlegenden Funktionen der Erfassung, Aufbereitung und Durchsuchbarmachung von Informationen aus dem Web zu erläutern. Diese grundlegende Funktionsweise gilt für alle Arten von Suchsystemen und wird für spezifische Anwendungszwecke angepasst. In diesem Abschnitt wird auf die Spezifika von Bildersuchmaschinen eingegangen, also solchen Suchmaschinen, mit denen nach Bildern, die öffentlich im Web verfügbar sind, gesucht werden kann.

Auch bei der Bildersuche besteht der Dreischritt von Crawling, Indexierung und Abgleich, wobei sich aufgrund der Verwendung von Bildern anstelle von Text Spezifika ergeben.

Wie oben erläutert, werden Bilder im Web durch ihre URL aufgefunden. Da Bilder selbst keine Links enthalten, kann ein alleiniges Crawlen von Bildern nicht erfolgen. Vielmehr müssen HTML-Dokumente gefunden werden, aus denen die URLs von Bildern extrahiert werden können. Dabei kann es sich um in den Text eingebettete Bilder handeln oder um Bilder, die aus dem HTML-Dokument heraus verlinkt, aber nicht im HMTL-Dokument angezeigt werden. Das Erfassen der URLs der Bilder kann also als Nebenprodukt des Crawlings von Textdokumenten betrachtet werden. Eine Suchmaschine, die sich allein auf Bilder beschränken würde, müsste ebenso wie eine Universalsuchmaschine alle HTML-Dokumente erfassen. Einzig der Speicherbedarf wäre geringer, da im Content Store die Textdokumente entfallen würden.

Web-Suchmaschinen verwenden eine Vielzahl von Crawlern zum Aufbau von speziellen Datenbeständen und weitere Zwecke (vgl. bspw. die Übersicht von Google, 2025), unter anderem spezialisierte Crawler für die Erfassung von Bildern (bspw. *Googlebot Image*). Die Funktion von Bildercrawlern ist, die vom allgemeinen Crawler übergebenen URLs von Bildern zu besuchen und diese Bilder herunterzuladen.

Die Vorverarbeitung der Bilder muss in anderer Form erfolgen als die der Textdokumente. Allerdings spielt auch in der Bildersuche Text eine wesentliche Rolle, und zwar zur Beschreibung des Inhalts des jeweiligen Bilds. Oft wird angenommen, Bildersuchmaschinen würden die Bilder visuell eingehend analysieren und mit Verfahren der Bilderkennung eine inhaltliche Beschreibung der Bilder erstellen. Dies ist jedoch nicht der Fall. Vielmehr wird der Text, der das jeweilige Bild umgibt, ausgewertet, da dieser häufig eine gute Beschreibung des Bildinhalts liefert. Beispielsweise steht häufig direkt unter dem Bild einer Person der Name der dargestellten Person. Die Auswertung von solchen Umgebungstexten kann mit weiteren Metadaten (wie dem Titel des Bilds, Alternativtext und Ankertexten) zu einer textuellen Beschreibung des Bilds kombiniert werden (Lewandowski, 2021, S. 54f.), was für die Beantwortung vieler Suchanfragen bereits ausreichend ist.

Weiterhin verwenden Bildersuchmaschinen sog. Low-Level-Features, also wesentliche Aspekte eines Bilds, wie zum Beispiel Konturen und Farbverteilungen. Diese können für die Ähnlichkeitssuche zwischen Bildern eingesetzt werden. Denkbar ist selbstverständlich auch die Verwendung von biometrischen Templates, diese werden aber bei den gängigen Bildersuchmaschinen nicht verwendet.

Die Repräsentation der Bilder in den Bildersuchmaschinen von Anbietern wie Google und Bing bestehen also aus einer Kombination einer textuellen Repräsentation mit Merkmalen, die mit Verfahren der Bilderkennung extrahiert wurden. Diese Kombination erlaubt sowohl die Suche mittels einer textuellen Eingabe als auch eine Ähnlichkeitssuche zwischen Bildern. Ebenso wie die textuellen Inhalte werden die Repräsentationen der Bilder in Indexen aufbereitet, um den Datenbestand effizient durchsuchbar zu machen.

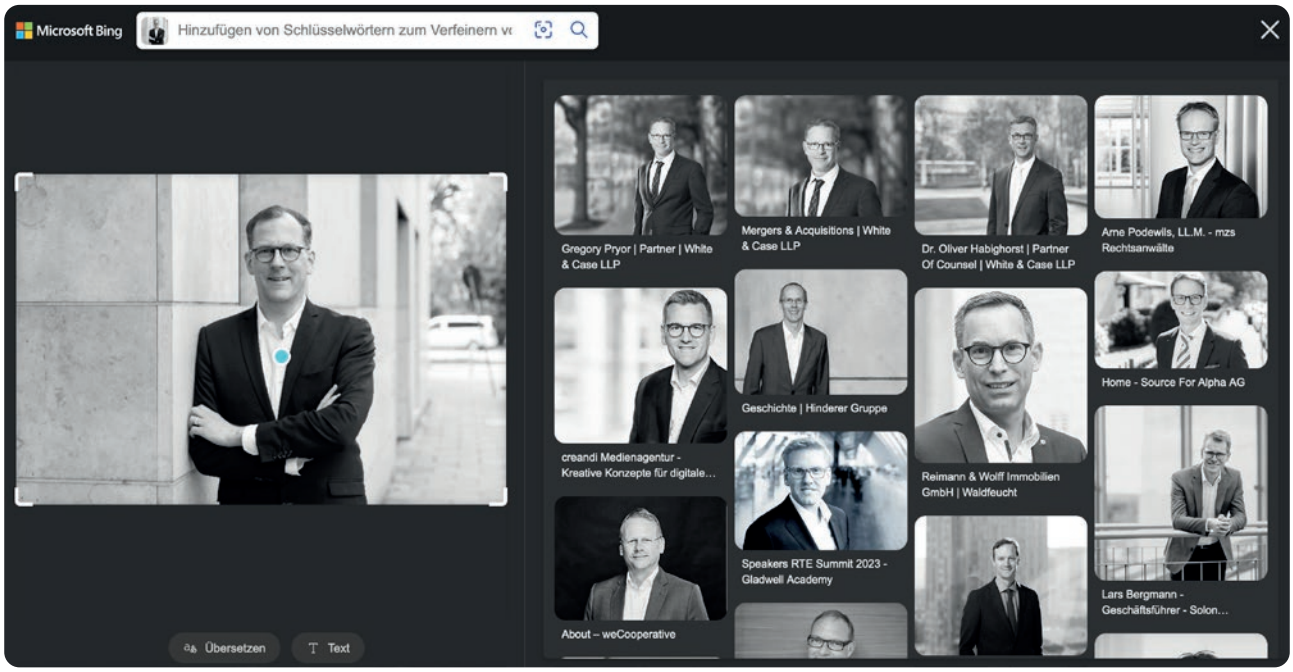


Abbildung 2: Ähnlichkeitssuche auf der Basis eines hochgeladenen Bilds (Bing; 19.7.2025)

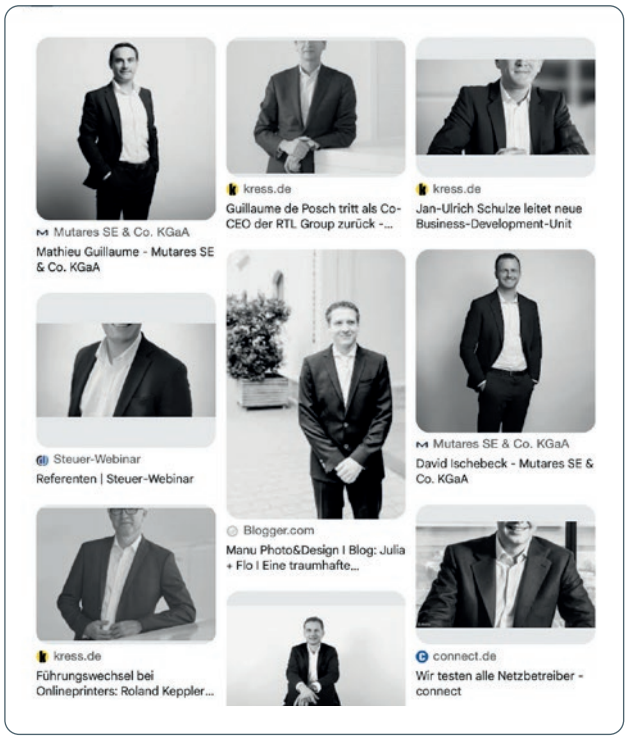


Abbildung 3: Ähnlichkeitssuche auf der Basis eines hochgeladenen Bilds (Google; 19.7.2025)

Bei der textuellen Suche in einer Bildersuchmaschine wird eine Suchanfrage wie gewohnt als textuelle Eingabe formuliert. Der Abgleich der Suchanfragen erfolgt mit den in den Repräsentationen der Bilder enthaltenen Texten. So können beispielsweise durch die Eingabe des Namens einer Person mit guter Treffgenauigkeit Bilder ebendieser Person gefunden werden. Die Grenzen des Ansatzes zeigen sich allerdings leicht, wenn entweder keine oder kaum Bilder einer Person im Datenbestand der Suchmaschinen vorhanden sind oder das Bild der Person in ein Textdokument mit Bildern mehrerer Personen eingebettet ist. In diesem Fall kann es zu einer Fehlzuordnung der Umgebungstexte kommen. Dies tritt häufig bei Mitarbeiterverzeichnissen auf Websites auf; im Suchergebnis werden dann für die Suche nach einer Person auch Bilder von Kollegen ausgegeben.

Bei der visuellen Suche wird ein Referenzbild, beispielsweise das einer Person, hochgeladen. Von diesem Bild wird eine Repräsentation erstellt, die mit den bei der Suchmaschine vorhandenen visuellen Repräsentationen der Bilder verglichen wird. Die gefundenen Bilder werden nach absteigender Ähnlichkeit angeordnet (Ranking). Die Ähnlichkeit wird über die Low-Level-Features gemessen. Wird eine vollständige bzw. nahezu vollständige Übereinstimmung gefunden, können weitere Informationen zu dem Bild über die textuelle Repräsentation des Bilds gewonnen werden, beispielsweise der Name der Person. Die initiale Suchanfrage kann dann mit diesen textuellen Informationen ergänzt und neu ausgeführt werden, bevor die endgültige Treffermenge erstellt wird.

Die Grenze dieses Ansatzes liegt darin, dass keine echten Ähnlichkeiten zwischen Personen gemessen werden, sondern nur visuelle Ähnlichkeiten zwischen Bildern. Dies lässt sich gut an dem Beispiel eines hochgeladenen Schwarz-Weiß-Bildes zeigen: Die Ähnlichkeitssuche findet nur weitere Schwarz-Weiß-Bilder, da die Farbverteilungen für die Berechnung der Ähnlichkeit verwendet werden (Abbildungen Abbildung 2 und Abbildung 3). Weiterhin wird im Beispiel deutlich, dass auf keinem der gefundenen Bilder die auf dem Originalbild gezeigte Person dargestellt ist, sondern vielmehr Personen, die in einer ähnlichen Ansicht gezeigt werden.

2.3.2 Biometrische Bildersuchmaschinen: Beispiel PimEyes

Um die Frage der „datenbankfreien" Suche besser in den Kontext des biometrischen Abgleichs von Bildern einordnen zu können, wird in diesem Abschnitt die Funktionsweise der Suchmaschine PimEyes beschrieben, die insbesondere durch den Fall des Auffindens von Bildern des ehemaligen RAF-Mitglieds Daniela Klette Bekanntheit erlangt hat (Killian, 2024; Schlereth, 2024). Ähnlich der Möglichkeit in der Google-Bildersuche, ein Bild hochzuladen und ähnliche Bilder zu finden, geht auch PimEyes vor, allerdings mit dem Unterschied, dass hier die biometrischen Repräsentationen (Templates) abgeglichen werden *(Image Search with PimEyes | PimEyes' Blog | PimEyes, o. J.)*.

Bei PimEyes wird ein Bild hochgeladen und mit dem Datenbestand abgeglichen. Um den Abgleich durchzuführen, hat auch PimEyes eine Datenbank angelegt und die Bilder vorverarbeitet. Die Daten werden ähnlich wie bei anderen Bildersuchmaschinen von öffentlich zugänglichen Websites zusammengeführt. Neben anderen Metadaten wird zu jedem Bild ein biometrisches Template angelegt und die URL gespeichert, auf der das Bild gefunden wird. PimEyes verwendet die Proportionen von Gesichtern und erstellt daraus die Templates; andere biometrische Merkmale wie Bewegungs- oder Sprechmuster könnten in ähnlichen Suchmaschinen analog verarbeitet werden.

Durch die Vorverarbeitung zu Templates kann nun der Abgleich Bild zu Bild aufgrund biometrischer Merkmale durchgeführt werden: Wird ein Bild hochgeladen, wird zu diesem Bild ein Template erstellt und mit den in der Datenbank vorhandenen Templates verglichen. Der Vergleich erfolgt also analog zu anderen Suchanwendungen zwischen Repräsentationen (Templates), nicht zwischen Bildern oder zwischen Texten und Bildern.

Die durch den Abgleich gefundenen Treffer werden auf Basis der angenommenen Ähnlichkeit in eine Reihenfolge (Ranking) gebracht. Auch hier wird kein absoluter Wert für die Übereinstimmung, sondern eine Wahrscheinlichkeit der Übereinstimmung angenommen. Der Bezug zwischen Bild und Personenname wird von PimEyes nicht hergestellt. Vielmehr zeigt das System ähnlich wie andere Suchmaschinen auch nur Vorschaubilder und verlinkt auf die URLs, auf denen das jeweilige Bild gefunden wurde. Dabei handelt es sich wiederum um das HTML-Dokument, in das das Bild eingebettet ist. Durch das Lesen des Umgebungstexts dort können Nutzende der Person auf dem Bild leicht einen Namen zuordnen. Während es bei allgemeinen Bildersuchmaschinen aufgrund des durch die Analyse der Low-Level-Features sehr beschränkten Abgleichs sehr unwahrscheinlich ist, dass eine Person, die einen anderen als den der abfragenden Person bekannten Namen verwendet, durch die Ähnlichkeitssuche gefunden wird, sind solche Treffer in PimEyes aufgrund des Abgleichs der Templates einfach möglich.

Durch den Abgleich der Templates ist PimEyes einfachen Bildersuchmaschinen wie der Google Bildersuche bei Personensuchen auf Basis eines Originalbilds weit überlegen. Die Überlegenheit in der Suche liegt darin, dass die biometrischen Gesichtsmerkmale für die Suche verwendet werden und nicht nur Low-Level-Features wie Formen und Farbverteilungen oder Umgebungstexte. Dadurch können Personen in unterschiedlichen Kontexten, in unterschiedlichen Positionen oder mit veränderten Merkmalen (bspw. unterschiedlicher Haarschnitt) gefunden werden. PimEyes ist also im Gegensatz zu anderen Bildersuchmaschinen eine Gesichtserkennungssoftware (Bovermann et al., 2024).

Aktuelle Zahlen zum Datenbestand von PimEyes sind nicht verfügbar. Im Jahr 2020 lag der Bestand allerdings schon bei 900 Millionen Bildern von Personen (https://web.archive.org/web/20200405180313/https://pimeyes.com/en/). Clearview, ein Anbieter ähnlicher Technologie, gibt an, eine Datenbank mit mehr als 60 Milliarden Bildern aufgebaut zu haben (https://web.archive.org/web/20250724005132/https://www.clearview.ai/).

2.3.3 Potenziell für den Abgleich geeignete, offene Bilderbestände

Neben der Möglichkeit, Bilder über eine eigene Erfassung im Web zu erschließen, besteht auch die Möglichkeit, auf gesonderte Kollektionen von Bildern, vor allem von Social-Media-Plattformen, zuzugreifen. Dabei muss unterschieden werden zwischen den öffentlich zugänglichen Bildern, die auch von den Crawlern der allgemeinen Suchmaschinen erfasst werden können, und den im geschlossenen Bereich des jeweiligen Systems vorhandenen Bildern, die nur eingeschränkten Nutzergruppen innerhalb des Systems zugänglich sind (bspw. Fotos, die nur „Freunden“ bei Facebook zugänglich sind). Aber auch, wenn man diese Unterscheidung trifft, lassen sich in Social-Media-Plattformen große Mengen öffentlich zugänglicher Bilder finden. Es ist möglich, diese Bilder mittels *Focused Crawling* (s. Abschnitt 4.2.2) zu erfassen oder, sofern der jeweilige Anbieter

eine solche Funktion anbietet, über ein Application Programming Interface (API), also eine Software-Schnittstelle, abzufragen.

Den Plattformen ist gemein, dass sie jeweils die von ihren Nutzern selbst hochgeladenen Bilder hosten, unabhängig davon, ob auf dem jeweiligen Bild nur die Person zu sehen ist, die das Bild hochgeladen hat oder auch weitere Personen. Auf den Plattformen hochgeladene Bilder können also auch Personen darstellen, die selbst nicht auf der Plattform angemeldet sind. Hinsichtlich der Indexierung der Daten verwenden die Plattformen die grundlegend gleichen Verfahren wie die Suchmaschinen. Damit bilden auch sie durchsuchbare Datenbanken von Bildern und anderen Inhalten.

Wichtige Plattformen, auf denen eine Vielzahl von Bildern von Personen zu finden sind, sind beispielsweise Facebook, Instagram, LinkedIn und X (vormals Twitter). Diese Plattformen veröffentlichen keine Zahlen zur Anzahl der gespeicherten Bilder, bekannt ist allerdings, dass bei Instagram bereits im Jahr 2016 jeden Tag 95 Millionen Bilder hochgeladen wurden (Reuters, 2016). Hier ist allerdings zu beachten, dass hier keine Unterscheidung zwischen Bildern, die Personen zeigen, und anderen Bildern getroffen wurde.

3 Datenbanken

Eine Datenbank ist eine Sammlung von Daten, die mittels eines Datenbankverwaltungssystems modifiziert und durchsucht werden kann (Kämper & Eickler, 2015, S. 21). Wenn die Rede von einer Datenbank ist, ist häufig die *Datenbasis* gemeint, d.h. der Bestand der in dem Datenbankverwaltungssystem erfassten Daten. In diesem Sinne werden die Bezeichnungen Datenbasis und Datenbank in diesem Gutachten synonym gebraucht. Ein Datensatz besteht aus der Beschreibung eines Informationsobjekts (bspw. Titel, URL, Dateigröße, Template) und ggf. dem Informationsobjekt selbst (bspw. Text oder Bild). Eine Datenbank besteht aus mehreren Datensätzen.

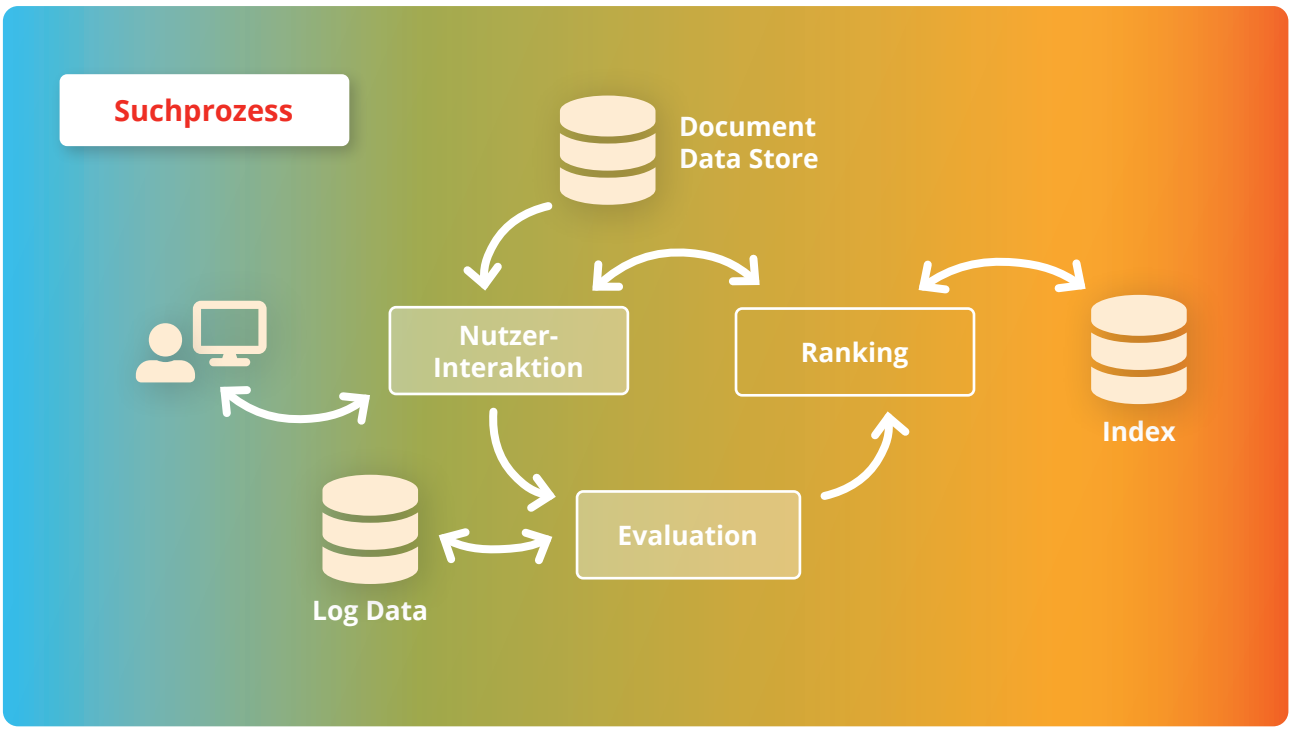


Abbildung 4: Suchprozess nach Croft, Metzler & Strohman, 2009, https://ciir.cs.umass.edu/downloads/SEIRiP.pdf

Die Datenbasis kann aus intern oder extern vorliegenden Daten aufgebaut werden. Wenn beispielsweise ein Presseverlag ein digital durchsuchbares Archiv der von ihm veröffentlichten Zeitungsartikel aufbaut, werden intern vorliegende Daten verwendet. Ebenso erstellt eine Social-Media-Plattform aus den von den Nutzenden hochgeladenen Bildern (intern vorliegende Daten) eine Datenbank. Im Gegensatz dazu greifen Suchsysteme, die öffentlich im Web verfügbare Daten erfassen, auf extern vorliegende Daten zu und bauen eine Datenbank aus Kopien der extern aufgefundenen Informationsobjekte auf. Gemeinsam ist beiden Arten von Systemen, dass sie die Daten in einer Datenbasis zusammenführen, die dann durchsuchbar gemacht wird.

Für die Bezeichnung einer Sammlung als Datenbank ist es unerheblich, ob die Daten dauerhaft oder temporär gespeichert werden. Gerade der sich beständig verändernde Datenbestand des Web erfordert, dass auch die Datenbanken der Suchsysteme, die diese Inhalte durchsuchbar machen, beständig aktualisiert werden. Dies ist unabhängig davon, ob ein Datenbanksystem nicht mehr im Web vorhandenen Informationsobjekte in der eigenen Datenbasis löscht oder beibehält, um so auch im Web nicht mehr vorhandene Inhalte weiterhin auffindbar zu halten.

Auch die Größe des Datenbestands ist kein kennzeichnendes Merkmal einer Datenbank. Sobald Datensätze angelegt werden, entsteht eine Datenbasis. Datenbankverwaltungssysteme sind darauf angelegt, auch große Datenbestände effizient durchsuchbar zu machen. Dafür ist wiederum eine umfassende Vorverarbeitung nötig, bei der Indexe angelegt werden (Kämper & Eickler, 2015). Eine Datenbank ohne Index ist möglich; es ergibt sich hier aber analog zu der in Abschnitt 2.1 beschriebenen „Live-Suche“ das Problem, dass bei jeder Suche *alle Datensätze* durchsucht werden müssen, was eine solche Datenbank nicht praktikabel macht. Ein Index als Verzeichnis einer Datenbank ist selbst wiederum eine Datenbank, die auf Einträge in einer Datenbank oder auf extern vorliegende Dokumente (bspw. über URLs) verweist.

4 Abgleich biometrischer Daten ohne Verwendung einer Datenbank

In diesem Abschnitt geht es schließlich um die Möglichkeiten des Abgleichs, für die kein Aufbau einer Datenbank erforderlich ist. Weiterhin wird erörtert, inwieweit die beschriebenen Lösungen in der Praxis umsetzbar sind.

4.1 Auf der Seite der anfragenden Stelle vorhandene Daten

Für die folgende Diskussion wird davon ausgegangen, dass die anfragende Stelle über alle oder einen Teil der folgenden Daten verfügt:

1. **Das Bild einer Person**
2. **Ein biometrisches Template, das aus dem Bild der Person erstellt wurde**
3. **Den Namen der abgebildeten Person**
4. **Weitere Metadaten zum Bild der Person, wie Zeitstempel und Aufnahmeort**

Die weiteren Metadaten werden im Folgenden nicht berücksichtigt, da sie nur bei der Verfeinerung des Abgleichs helfen können, selbst aber keinen grundlegenden Einfluss auf den Prozess des Abgleichs haben. Ebenso bleibt der Fall unberücksichtigt, dass die anfragende Stelle über mehrere Bilder der gleichen Person verfügt. Dies kann zwar die Qualität der Ergebnisse verbessern, verändert aber nicht den grundlegenden Prozess der Abfrage.

Ohne eine Datenbank aufzubauen, kann kein One-to-many-Abgleich durchgeführt werden, sondern nur ein One-to-One-Vergleich. Bei einem One-to-One-Vergleich werden zwei Bilder daraufhin überprüft, ob sie dieselbe Person zeigen. Um das Erstellen einer

Datenbank zu vermeiden, kann das Originalbild mit anderen Bildern nur in einer Vielzahl von One-to-One-Abgleichen verglichen werden. Sobald hingegen ein One-to-Many-Abgleich erfolgen soll, muss eine Datenbank erstellt werden, mit der dann das Originalbild abgeglichen wird.

4.2 Live-Suche im Web

In den vorangegangenen Abschnitten wurde gezeigt, dass erst die Sammlung und Vorverarbeitung der Daten eine effiziente Suche ermöglicht. Dennoch soll in den folgenden Abschnitten gezeigt werden, wie eine „Live-Suche" im Web erfolgen könnte und warum sie zwar in der Theorie möglich, allerdings nicht umsetzbar ist. Dazu werden verschiedene Varianten der Live-Suche diskutiert.

4.2.1 Web-Crawling mit Einzelabgleich aller Bilder

Das Vorgehen des Web-Crawlings mit einem Einzelabgleich aller Bilder entspricht der in Abschnitt 2 beschriebenen „Live-Suche", bei der erst in dem Moment, in dem eine Suchanfrage vorliegt, mit der Erfassung der Web-Inhalte gestartet wird.

Im ersten Schritt wird ein biometrisches Template des Originalbilds angefertigt. Das Crawling erfolgt dann analog zu dem in Abschnitt 2.2.1 beschriebenen Vorgehen, allerdings wird nun die Besonderheit berücksichtigt, dass keine Datenbank aufgebaut werden soll. Der Crawler arbeitet wie beschrieben die Web-Dokumente nacheinander ab und folgt den in den Dokumenten gefundenen Links. Sobald die URL eines Bilds gefunden wurde, wird das Bild heruntergeladen. Zu diesem Bild wird nun ein biometrisches Template erstellt und mit dem Template des Originalbilds verglichen. Wird keine oder eine nur unzureichende Übereinstimmung festgestellt, wird das Bild gelöscht und der Crawlingprozess fortgesetzt. Wenn eine hinreichende Übereinstimmung gefunden wird, wird das Bild mit seiner URL und der URL des HTML-Dokuments, in dem das Bild enthalten ist, gespeichert.

Dieser Crawlingprozess würde potenziell das Web vollständig durchgehen und damit das Originalbild mit allen im Web vorhandenen Bildern abgleichen. Natürlich kann bereits vorher ein Match gefunden werden; um sicherzustellen, dass alle Vorkommnisse abgedeckt sind, muss der Crawl aber vollständig durchgeführt werden.

Dieses Vorgehen ist in der Theorie möglich, in der Praxis allerdings nicht durchführbar. Weiter oben wurden bereits die Dimensionen des Web und die Menge der dort vorliegenden Dokumente (inkl. Bildern) beschrieben. Auf dieser Basis kann selbst bei einer schnellen Verarbeitung der Dokumente die Aufgabe nicht in einer in irgendeiner Weise vertretbaren Zeit erledigt werden. Während im konventionellen Crawling ohne Parallelisierung der größte Zeitverzug durch das Warten auf die Antworten der angefragten Server entsteht, kommen bei dem hier beschriebenen Vorgehen noch folgende Faktoren des Zeitverzugs hinzu: (1) Zeit für das Herunterladen des Bilds, (2) Zeit für die Erstellung des Templates, (3) Zeit für den Abgleich mit dem Template des Originalbilds, (4) Zeit für das Löschen des heruntergeladenen Bilds.

Während die Zeiten für jeden dieser Schritte gering sein mögen, ist zu berücksichtigen, dass im Crawling ein möglichst vollständiges Abbild des Web mit allen enthaltenen Bildern geschaffen werden muss. Es wird damit ersichtlich, dass eine „Live-Suche" zwar theoretisch denkbar, in der Praxis aber nicht durchführbar ist. Der Zeitaufwand wäre im Hinblick auf die Größe des Web so groß, dass Ergebnisse auf eine Suchanfrage hin erst in Monaten oder gar Jahren erzielt werden könnten. Weiterhin hätte sich in dieser Zeit der Datenbestand des Web so stark verändert, dass die Suche von vorn beginnen müsste.

Bei dieser Betrachtung wurden bislang Fragen der Crawling-Etikette (s. Abschnitt 2.2.1) nicht berücksichtigt. Das beschriebene Vorgehen würde die im Web vorhandenen Server unnötig und übermäßig belasten und es ist davon auszugehen, dass, sobald ein solches Vorgehen bekannt würde, viele Website-Betreiber den entsprechenden Crawler von ihren Websites aussperren würden.

Es ist auch zu betonen, dass mit diesem Vorgehen auch keine Bilder gefunden werden können, die aktuell nicht (mehr) im Web vorhanden sind. Das System wäre also auch in der Hinsicht ineffizient, dass eine Person, die nicht wünscht, dass ein Bild, das sie selbst ins Netz gestellt hat, gefunden wird, dieses Bild einfach nur löschen müsste. Im Gegensatz kann eine speziell aufgebaute Datenbank, die Kopien der gefundenen Bilder erstellt, auch solche an der Originaladresse gelöschten Bilder finden.

4.2.2 Focused Crawling

Focused Crawling bezeichnet das Crawling eines eingeschränkten Bereichs des Web. Eine solche Einschränkung kann thematisch oder anhand formaler Merkmale (bspw. nur Nachrichtenseiten) erfolgen. Die einfachste Form der Einschränkung ist die manuelle Beschränkung auf eine oder mehrere Websites.

Das Vorgehen des Bildabgleichs mit einer Live-Suche mit Focused Crawling entspricht dem im letzten Abschnitt beschriebenen Vorgehen, allerdings mit der Beschränkung auf bestimmte Websites. Um ein sinnvolles System aufzubauen, müssten hier allerdings Websites ausgewählt werden, die besonders viele Bilder enthalten. Damit würde zwar der beschriebene Zeitbedarf reduziert, allerdings zulasten der Qualität der Ergebnisse, d.h. der Wahrscheinlichkeit, dass ein passendes Bild gefunden wird. Weiterhin ist zu berücksichtigen, dass auch einzelne relevante Websites, wie die der Social-Media-Plattformen, Milliarden von Bildern enthalten, was auch bei einer eingeschränkten Leistung zu einem übermäßigen Zeitbedarf führt.

4.3 Zugriff über Application Programming Interfaces (APIs)

Application Programming Interfaces (APIs) erlauben den strukturierten Zugriff auf die (auch über das Web-Interface öffentlich zugänglichen) Inhalte eines Anbieters. Allerdings ist zu beachten, dass solche API-Zugänge nur von wenigen Plattformen angeboten werden und Restriktionen hinsichtlich der Anfragen und der Verwendung der Daten beinhalten. Sie können damit das Crawling nicht ersetzen. Weiterhin wird im Folgenden davon ausgegangen, dass die jeweilige API Zugriff auf alle Bilder erlaubt, d.h. nacheinander jedes Bild aus dem Datenbestand ohne die Formulierung einer Suchanfrage abgefragt werden kann. Das im Folgenden beschriebenen Vorgehen bezieht sich also auf eine hypothetische Situation.

Das Vorgehen wäre prinzipiell das gleiche wie beim beschriebenen Crawling-Ansatz, außer dass das Auffinden der Bilder entfallen würde. Stattdessen würde jedes einzelne Bild aus dem Datenbestand abgerufen, ein Template erstellt, dieses mit dem Template des Originalbilds verglichen und das heruntergeladene Bild gelöscht. Dadurch ergäben sich Zeitvorteile gegenüber dem Crawling, allerdings könnten jeweils nur die Bilder eines Anbieters abgefragt werden (sofern dieser Anbieter eine solche umfassende Abfrage überhaupt erlauben würde). Auch wenn die Daten mehrerer Anbieter abgefragt werden könnten, ergäbe sich gegenüber dem angestrebten Abgleich *mit allen Bildern aus dem Web* eine so erhebliche Einschränkung, dass die Lösung auch in dieser Hinsicht nicht praktikabel wäre.

4.4 Scraping von Bildersuchmaschinen

Zuletzt soll die Möglichkeit diskutiert werden, bestehende Bildersuchmaschinen wie die Google Bildersuche auszunutzen, um nicht selbst eine Datenbank anlegen zu müssen, gleichzeitig aber von dem großen, aus verteilten Quellen zusammengestellten Datenbestand zu profitieren. Dabei muss es sich um Suchmaschinen handeln, die zu den Bildern nicht selbst biometrische Templates anlegen (wie dies etwa bei PimEyes geschieht).

Um auf bestehende Bildersuchmaschinen automatisiert zugreifen zu können, müssen Nutzerinteraktionen mit der Suchmaschine simuliert werden. Ein automatisiertes System stellt Suchanfragen an eine Bildersuchmaschine und erfasst die Suchergebnisse mittels *Screen Scraping*. Dabei handelt es sich um ein Verfahren, mit dem Inhalte aus Webseiten

ausgelesen und gespeichert werden können, beispielsweise aus Suchergebnisseiten von Suchmaschinen. Nicht berücksichtigt werden an dieser Stelle rechtliche Beschränkungen, die diesen Ansatz einschränken.

Das Vorgehen wäre in diesem Fall wie folgt: (1) Zu einem Originalbild werden entweder textuelle Suchanfragen generiert, das Bild als Suchanfrage verwendet (vgl. Abschnitt 2.3.1) oder Kombinationen aus Text und Bild erstellt. (2) Die Suchanfragen werden automatisiert nacheinander an die Bildersuchmaschine geschickt; die Suchergebnisseiten werden gescrapt. Beim Scrapen wird nacheinander jedes einzelne Bild heruntergeladen, ein Template des Bilds erstellt, mit dem Template des Originalbilds verglichen, das heruntergeladene Bild gelöscht. Dieses Vorgehen wird für jedes Bild wiederholt (analog dem in Abschnitt 4.2.1 beschriebenen Vorgehen).

Theoretisch kann mit diesem Vorgehen sowohl das Problem der Erstellung einer Datenbank als auch das Problem der nur theoretisch vorhandenen Möglichkeit des Abgleichs mit allen im Web vorhandenen Bildern gelöst werden. Um dies zu erreichen, müssten allerdings Suchanfragen so gestellt werden können, dass sie eine Vorauswahl passender Bilder identifizieren könnten, mit denen sinnvoll der Abgleich der Templates durchgeführt werden kann. Dies ist in der Praxis allerdings nicht möglich.

Zunächst soll das Vorgehen mit textuellen Suchanfragen betrachtet werden. Natürlich kann mit dem Namen der gesuchten Person gesucht werden. Der Abgleich würde dann mit den Umgebungstexten (vgl. Abschnitt 2.3.1) erfolgen. Dieser Abgleich wäre trivial und nicht hilfreich, da das Problem ja gerade darin liegt, Bilder einer Person zu identifizieren, ohne dass diese Person im Zusammenhang mit dem Bild mit ihrem realen Namen benannt ist.

Weiterhin kann das Originalbild in die Bildersuchmaschine hochgeladen werden, um optisch ähnliche Bilder aufgrund der Low-Level-Features zu finden. Wie in Abschnitt 2.3.1 gezeigt wurde, führt dies allerdings nur mit einer verschwindend geringen Wahrscheinlichkeit zu relevanten Treffern, bei denen die Chance besteht, dass sie in dem folgenden Abgleich der Templates zu einem Erfolg führen. Dies gilt auch, wenn eine Vielzahl von Treffern aus der Bildersuchmaschine gescrapt werden könnte, was allerdings schon deshalb nicht möglich ist, weil Anbieter von Bildersuchmaschinen die Zahl der Bilder, die pro Suchanfrage angezeigt werden, stark einschränken – auch bei einer manuellen Suche.

Zuletzt kann eine Kombination von Bild und Text für die Generierung von Suchanfragen verwendet werden. Dabei wird das Bild in die Bildersuchmaschine hochgeladen und automatisiert mit (nicht notwendigerweise sinntragenden) Suchwörtern kombiniert. Dies bedeutet, dass eine Vielzahl von Suchanfragen generiert wird, die jeweils aus dem gleichen Bild, allerdings kombiniert mit jeweils anderen Suchwörtern, bestehen. Das Ziel ist es, eine große Anzahl von Bildern zu erreichen, die optische Ähnlichkeit zu dem Originalbild haben, und gleichzeitig die von den Bildersuchmaschinen gesetzten Beschränkungen der Treffermengen zu überwinden. Ein solches Vorgehen kann zu einer großen Zahl optisch ähnlicher Bilder führen; seine Beschränkung liegt aber wiederum im Kriterium der optischen Ähnlichkeit. Damit kann mit diesem Verfahren nur mit einer verschwindend geringen Wahrscheinlichkeit ein Bild der Person ohne die Nennung ihres Namens und ohne sehr starke Ähnlichkeit der Bilder hinsichtlich der Farbgebung und Pose erzielt werden.

5 Fazit

Die Analyse hat gezeigt, dass Bilder aus dem Web ohne die Erstellung einer Datenbank nicht sinnvoll durchsuchbar gemacht werden können. Die Darstellung theoretischer Ansätze, die dies ermöglichen könnten, hat gezeigt, dass diese in der Praxis allesamt scheitern müssen.

6 Literaturverzeichnis

Bovermann, M., Fink, J., & Mutter, J. (2024). PimEyes User auf den Spuren der RAF. *Verfassungsblog*. https://doi.org/10.59704/ee5e12eaf02c1341

Croft, W. B., Metzler, D., & Strohman, T. (2009). Search Engines: *Information retrieval in practice*. Pearson. https://ciir.cs.umass.edu/downloads/SEIRiP.pdf

Google. (2025). *Google's common crawlers | Google Search Central | Documentation | Google for Developers*. https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers

Image search with PimEyes | PimEyes' Blog | PimEyes. (o. J.). Abgerufen 5. Juli 2025, von https://pimeyes.com/en/blog/image-search-with-pimeyes-how-to-reverse-image-search

Kämper, A., & Eickler, A. (2015). *Datenbanksysteme: Eine Einführung* (10. Aufl.). De Gruyter Oldenbourg.

Killian, N. (2024, März 22). Wie viel Überwachung ist zu viel? *Süddeutsche Zeitung*.

Lewandowski, D. (2021). *Suchmaschinen verstehen* (3. Aufl.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-63191-1

Reuters. (2016, Juni 21). Instagram's user base grows to more than 500 million. *Reuters*. https://www.reuters.com/article/technology/instagrams-user-base-grows-to-more-than-500-million-idUSKCN0Z71LN/

Schlereth, P. (2024, März 2). Ein Journalist fand die RAF-Terroristin in 30 Minuten. *Frankfurter Allgemeine Zeitung*.

Schwartz, B. (2016). *Google's search knows about over 130 trillion pages*. Search Engine Land. http://searchengineland.com/googles-search-indexes-hits-130-trillion-pages-documents-263378

Shepard, C. (2023, November 27). Google's Index Size Revealed: 400 Billion Docs. *Zyppy Marketing*. https://zyppy.com/seo/google-index-size/

Smith, N., Manager, P., & Images, G. (2010). Ooh! Ahh! Google Images presents a nicer way to surf the visual web. *Official Google Blog*. https://googleblog.blogspot.com/2010/07/ooh-ahh-google-images-presents-nicer.html

Teevan, J., Collins-Thompson, K., White, R. W., & Dumais, S. (2014). Slow Search: Seeking to enrich the search experience by allowing for extra time and alternate resources. *Communications of the ACM*, 57(8), 36–38.

7 Kurzbiografie des Autors

Dirk Lewandowski ist Professor für Information Research & Information Retrieval an der Hochschule für Angewandte Wissenschaften Hamburg. Von 2022 bis 2024 war er Vertretungsprofessor für Data Science an der Universität Duisburg-Essen, Fakultät für Informatik. Im Sommersemester 2021 war er als Visiting Research Fellow am Pufendorf Institute for Advanced Studies an der Universität Lund, Schweden, tätig. Seine Forschungsinteressen sind Web Information Retrieval, Data Science, Qualitätsfaktoren von Suchmaschinen, das Recherchevehalten der Suchmaschinen-Nutzer sowie die gesellschaftlichen Auswirkungen des Umgangs mit den Web-Suchmaschinen.

Zu seinen Veröffentlichen gehören neben den Büchern „Suchmaschinen verstehen" (3. Auflage 2021), „Web Information Retrieval", „Handbuch Internet-Suchmaschinen" (drei Bände) und „Web Search Engine Research" zahlreiche Aufsätze, die in deutschen und internationalen Fachpublikationen veröffentlicht wurden. Als Experte war er unter anderem für den High Court of Justice (Großbritannien) und den Deutschen Bundestag tätig. Er ist einer der Distinguished Members der Association for Information Science & Technology (ASIST) Von 2013 bis 2020 war Dirk Lewandowski Editor-in-Chief der Zeitschrift Aslib Journal of Information Management. Er ist Mitglied in den Editorial Boards mehrerer internationaler Zeitschriften, darunter Journal of the Association for Information Science & Technology (JASIST).

Ausbildung

Studium des Bibliothekswesens (Dipl.-Bibliothekar (FH), Hochschule für Bibliotheks- und Informationswissenschaft, 1997), der Philosophie, Informationswissenschaft und Medienwissenschaft (M.A., Heinrich-Heine-Universität Düsseldorf, 2002), Promotion in Informationswissenschaft (Dr. phil., Heinrich-Heine-Universität Düsseldorf, 2005), Psychologie (B.Sc., Fernuniversität in Hagen, 2020).

Beruf

Ministerium für Wirtschaft und Mittelstand, Energie und Verkehr des Landes Nordrhein-Westfalen (1997-2001, Bibliothekar und Informationsvermittler), NRW Medien GmbH (2001-2004, Projektleiter Research), Heinrich-Heine-Universität Düsseldorf, Abteilung Informationswissenschaft (2004), Institut für Forschungsinformation und Qualitätssicherung Bonn (2006), Freiberufler (2004-2007).

Dirk Lewandowski
Professor für Information Research & Information Retrieval an der Hochschule für Angewandte Wissenschaften Hamburg

/ Impressum

ALGORITHM WATCH

AW AlgorithmWatch gGmbH
Boyenstraße 41
10115 Berlin
www.algorithmwatch.org
policy@algorithmwatch.org

Autor
Prof. Dr. Dirk Lewandowski
Hochschule für Angewandte Wissenschaften Hamburg
dirk.lewandowski@haw-hamburg.de

Layout & Illustration
Beate Autering, Benito Felkel

Die Publikation steht unter der Lizenz Creative Commons Namensnennung 4.0 International (CC BY 4.0).
https://creativecommons.org/licenses/by/4.0/deed.de