Bildanalyse Software

Betreuer: Matthias Bachmann

Studierende: Roger Bollmann

Datum: September 2015

Inhaltsverzeichnis

[Aufgabenstellung 4](#_Toc415143949)

[Thema 4](#_Toc415143950)

[Mitwirkende Personen 4](#_Toc415143951)

[Ausgangslage 4](#_Toc415143952)

[Ziel der Arbeit 4](#_Toc415143953)

[Aufgabenstellung 4](#_Toc415143954)

[Erwartete Resultate 4](#_Toc415143955)

[Recherche 5](#_Toc415143956)

[Markanalyse 5](#_Toc415143957)

[ExperVision 5](#_Toc415143958)

[Maestro Recognition Server 5](#_Toc415143959)

[MaxxVault LLC 6](#_Toc415143960)

[Fazit 6](#_Toc415143961)

[Methoden zum Abfangen von Bilder 6](#_Toc415143962)

[Modul 6](#_Toc415143963)

[Advances Logging 6](#_Toc415143964)

[Filter 6](#_Toc415143965)

[Übertragungsmethoden 6](#_Toc415143966)

[Windows Service 6](#_Toc415143967)

[API 7](#_Toc415143968)

[REST 7](#_Toc415143969)

[Websocket 7](#_Toc415143970)

[Sicherheit 7](#_Toc415143971)

[Push oder Pull? 7](#_Toc415143972)

[Textanalyse Software 7](#_Toc415143973)

[Anforderungen 7](#_Toc415143974)

[Konzept 8](#_Toc415143975)

[Proof of Concept „PoC“ 8](#_Toc415143976)

[Sender 8](#_Toc415143977)

[Empfänger 8](#_Toc415143978)

[Translator 8](#_Toc415143979)

[Testing 8](#_Toc415143980)

[Unit Test 8](#_Toc415143981)

[User Akzeptanz Tests 8](#_Toc415143982)

[Fazit 8](#_Toc415143983)

# Aufgabenstellung

## Thema

Abfangen von Bilder von einem Webserver und umwandeln in Text zur Analyse.

## Mitwirkende Personen

Studierender Roger Bollmann

Betreuungsperson Matthias Bachmann

## Ausgangslage

Als Reaktion auf den Diebstahl von zahlreichen Kundendaten, sowie des anschliessenden Verkaufs von illegalen CDs, hat die Eidgenössische Finanzmarktaufsicht (FINMA) alle Schweizer Finanzinstitute aufgefordert, den Zugriff auf Kundendaten verstärkt zu überwachen. Das Unternehmen für das ich arbeite, integriert nun eine Überwachungssoftware, die den HTTP-Traffic nach Kundendaten absucht. Die Software ermöglicht jedoch lediglich eine systematische Analyse von Text. Bilder, die ebenfalls Kundendaten enthalten könnten, werden ignoriert. Grund dafür ist, dass auf dem Markt keine wirklich effizienten Bildanalyse-Softwares angeboten werden.

## Ziel der Arbeit

Das Ziel der Arbeit ist ein Programm zu entwickeln, welches die versendeten Bilder von einem Webserver in Text umwandelt, um eine Analyse durch die Überwachungssoftware zu ermöglichen. Das Programm sollte aus zwei Teilen bestehen, einem Sender und einem Empfänger. Der Sender wird zuständig sein für das Versenden der Bilder, wobei der Empfänger die Bilder empfangen und in Text umwandeln soll.

## Aufgabenstellung

1. Recherche
   1. Analyse von verschiedenen Methoden zum Abfangen von Bilder auf einem Webserver
   2. Analyse von verschiedenen Übertragungsmethoden
2. Anforderungen ermitteln und dokumentieren
3. Erstellen eines Konzepts
   1. Vergleichen von verschiedenen Textanalytik-Applikationen
4. Erstellen eines Proof of Concepts
   1. Implementierung eines Senders, welcher die Bilder abfängt und an den Empfänger weiterleitet
   2. Implementierung eines Empfängers, welcher die Bilder empfängt und in Text umwandelt
5. 5.Fazit

## Erwartete Resultate

1. Durchführen einer Recherche
2. Definition einer Anforderungsanalyse
3. Ausarbeitung eines Konzept
4. Durchführen eines Proof of Concept „PoC“
5. Ausführung von Tests
6. Persönliches Fazit

# Recherche

## Markanalyse

### ExperVision

<http://www.expervision.com/ocr-software>

Negative

* + - 5200$ kostet die Lizenz für OpenRTK (SDK) Developer Tool, damit man selber entwickeln kann
    - Muss alles manuell eingestellt werden, Directory welche gescannt werden sollen. Vor allem für Dokumentenablagen und Emails zu empfehlen
    - Keine out of the Box Automatisierung, man muss sehr viel selber Programmieren, damit man das gewünscht Resultat erreicht.
    - Ein reines OCR Tool
    - Es können nur Bilder umgewandlet werden, jedoch kann nicht nachvollzogen werden, wer dieses Bild gesehen oder angeschaut hat
    - Keine Lösung zum anzeigen welcher User was gesehen hat
    - Keine Lösung für veschickte Bilder von Webserver nur Client oder Netzlauftwerk Scanning.

Positiv

* + - Unterstützt sehr viele Sprachen
    - Sehr viel Erfahrung (Gründung 1987)
    - Verfügt über viele Funktionen (Client/Server Scanning, Cloud-Service)
    - Unterstütz auch Mobile Bilder Umwandlung
    - Verfügt über ein API

### Maestro Recognition Server

<https://www.cvisiontech.com/products/general/maestro-recognition-server.html?lang=eng>

Negativ

* + - Keine Lösung zum anzeigen welcher User was gesehen hat
    - Keine Lösung für veschickte Bilder von Webserver nur Client oder Netzlauftwerk Scanning.

Positiv

* + - Ist sehr schnell auch bei grossen Datenmengen
    - Sobald ein Bild in einem Verzeichnis auftaucht, wird es automatisiert umgewandelt
    - Vorallem zum Umwandlen von PDF’s in suchbare Texte.

### MaxxVault LLC

<http://www.primerecognition.com/maxxvault.htm>

Architektur: <http://www.primerecognition.com/prime_network.htm>

Negativ

* + - Keine Lösung zum anzeigen welcher User was gesehen hat
    - Keine Lösung für veschickte Bilder von Webserver nur Client oder Netzlauftwerk Scanning.
    - Lizenz startet bei 4500$ für Job Server und zusätzlich noch pro Verifikations Server

Positiv

* + - Speichert die Daten auf einem sicheren und einfach zugreifbaren System.
    - Um möglichst viele Fehler auszuschliessen werden die Bilder durch mehrere ORC Engines umgewandlet.
    - Die Software hat ein Failover, falls es ein Fehler gibt bei einer OCR Engine, wird es ignoriert und für das nächste Bild vorbereitet.
    - Unterstütz 11 Sprachen
    - Kann durch Hinzufügen von Verfikations Server performanter Bilder in Text umwandeln (Cluster System)

### Fazit

Momentan sieht es so aus, dass es sehr gute Lösungen im Bereich ORC (Optical Character Recognition) gibt, jedoch keine welche meiner Anforderung entspricht. Bei allen Produkten gibt es einen Viewer, welche alle hinzugefügten Netzlaufwerke scannen und automatisiert prozessieren kann. Die hat zur Folge, dass bei einem neu hinzugefügten Netzlaufwerk, die Konfiguration angepasst werden muss. Bei grösseren Unternehmen möchte man viele Änderungen in der Produktion jedoch vermeiden.

## Methoden zum Abfangen von Bilder

### Modul

### Advances Logging

### Filter

Was ist mit Images in PDF Files?

<http://www.foolabs.com/xpdf/README>

<http://www.foolabs.com/xpdf/download.html>

C:\Users\Roger\Documents\GitHub\Bildanalyse\Tools\xpdfbin-win-3.04\bin32>pdfimages -j C:\Users\Roger\Documents\GitHub\Bildanalyse\Semesterarbeit\Test.pdf C:\Users\Roger\Documents\GitHub\Bildanalyse\Semesterarbeit\

Using xpdf in Visual Studio

<http://www.glyphandcog.com/manuals/XpdfPrintCOM/using.html>

## Übertragungsmethoden

### Windows Service

Vorteil

Nachteil

### API

Vorteil

Nachteil

### REST

Vorteil

Nachteil

### Websocket

Vorteil

Nachteil

### Sicherheit

Wie hohe Sicherheit muss implementiert werden?

### Push oder Pull?

## Textanalyse Software

OCR Software

# 

# Anforderungen

Use Cases

1. Use Case

Ein User sieht sich auf einer Webseite einige Bilder an, welche potentiell Kundendaten enthalten kann. Der Sender bekommt das mit und schickt das Bild, zur Überprüfung ob es Kundendaten enthält, an den Empfänger weiter. Der Empfänger wandelt das Bild in Text um und kann den Text zur Analyse weiterleiten.

1. Webserver schreibt ins Log, dass ein Bild verschickt wird
2. Der Sender hat ein Listener auf dem Logfile. Sobald ein neuer Eintrag hinzugefügt wird, liest er daraus die nötigen Informationen.
3. Der Sender kopiert das Bild in einen neuen Ordner, um es danach an den Empfänger weiterzuleiten.
4. Der Sender sendet das Bild über die ausgewählte Übertragungsmethode an den Empfänger
5. Der Empfänger bekommt das Bild und legt es an einen bestimmten Ort ab
6. Der Translator nimmt das Bild auf und wandelt es in Text um und legt der Text in einem bestimmten Ordner ab.

Webserver

Der Webserver muss über eine erweitertes Logging verfügen. Entweder wird das mit dem Package direkt mitinstalliert oder es muss zur Verfügung gestellt werden. Der Webserver muss so eingestellt werden dass er in ein separates Logfile mit folgenden Informationen liefert:

* Abfangen von Bilder welche verschickt werden
* Mime Type
* Status Code
* Usernamen (LDAP/SSO)
* Applikationsnamen optional
* Servernamen optional

Split muss ein Leerzeichen “ “ sein.

Sender

Der Sender hat zwei Funktionen. Einerseits muss er das Bild abfangen, welches vom Webserver an den User geschickt wird, anderseits muss er das Bild weiterleiten an den Empfänger.

1. Listener auf Logfile
2. Bild kopieren in ein bestimmtes Verzeichnis
3. Übertragung des Bildes inklusiv löschen des kopierten Bildes

Speicherplatz??

Emergency Stop??

Wie überträgt man die Bilder mit zusätzlichen Informationen? 🡪 Manipulieren des RequestHeaders

Übertragung zum Empfänger

Empfänger muss wissen welcher Server nicht mehr senden kann 🡪 DB [Servernamen, Status] /Front End zum Anzeigen der Daten.

Lokal gespeichert werden 🡪 Speicherplatz

Failover 🡪 Falls der Sender nicht senden kann 🡪 Loadbalancer/High Availability/Cloud Service

Emergency Stop 🡪 Automatisch (Memory/Verfügbarer Speicherplatz)

Emergency Stop 🡪 Manuell (Befehl zum stoppen evtl. Service einrichten)

Verschlüsslung?

# Konzept



+ Webfrontend

+ Verschlüsselte Übertragung

+ Möglichkeit zum Weiterleiten (Konfigurationsfile)

# Proof of Concept „PoC“

## Sender

## Empfänger

## Translator

# Testing

## Unit Test

## User Akzeptanz Tests

# Fazit