

Detekce formulářových polí ve skenovaných dokumentech

Michal Moravec
xmorav28@stud.fit.vutbr.cz

Vedoucí: Ing. Vítězslav Beran, Ph.D.

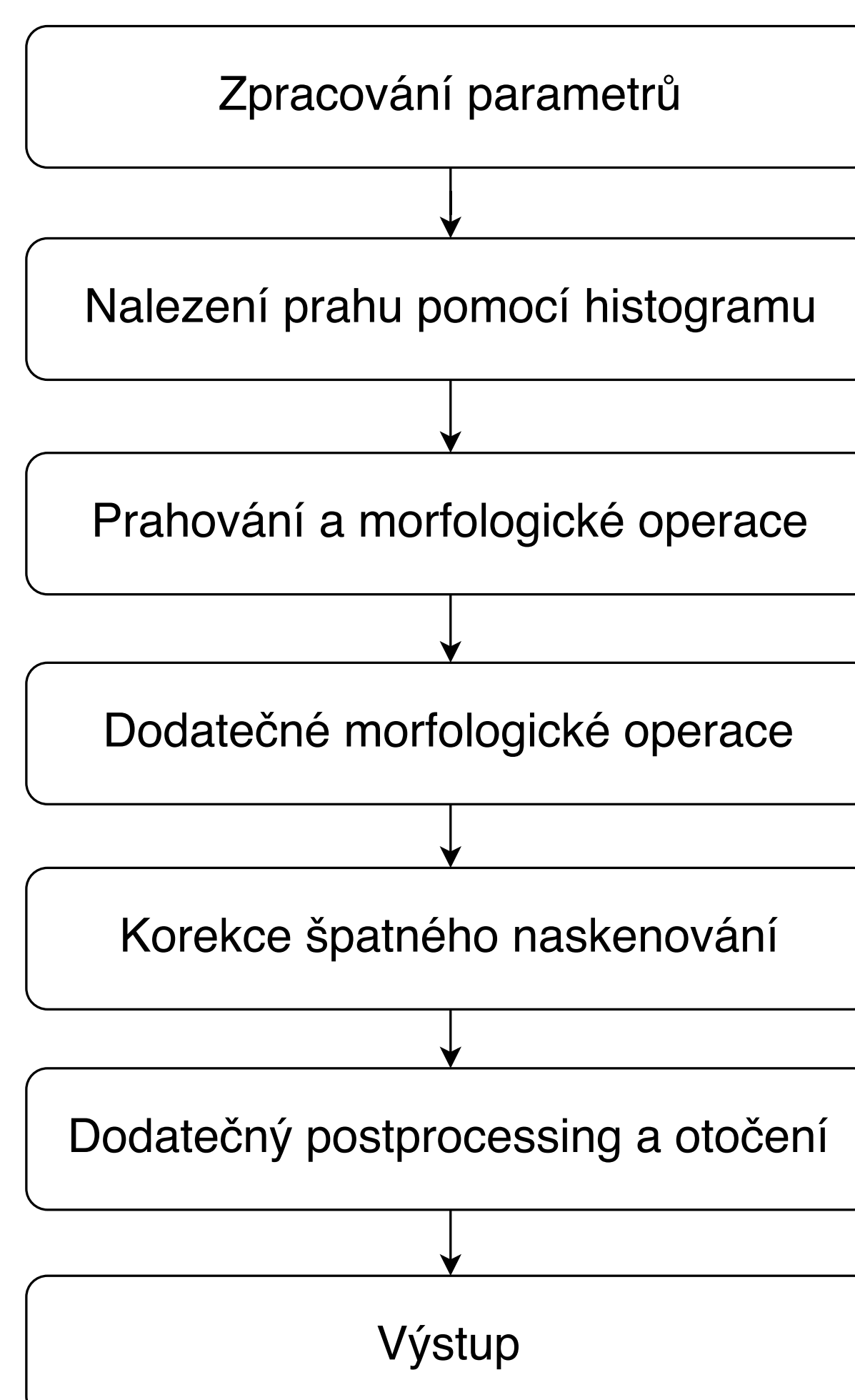
Abstrakt

Cílem práce bylo navrhnout algoritmus, který bude schopen z karty zdravotní pojišťovny vyseparovat pouze textová pole, která se budou dát dále použít v libovolném softwaru na převedení obrázku na text. Program by měl počítat se špatně naskenovanými a libovolně otočenými kartami. Celkový projekt je dělán jako zakázka pro firmu Medingo, která chce algoritmus zakomponovat do svého stávajícího systému. Co se týče výsledků, tak algoritmus dokáže detekovat a vyseparovat textová pole s velmi vysokou pravděpodobností.

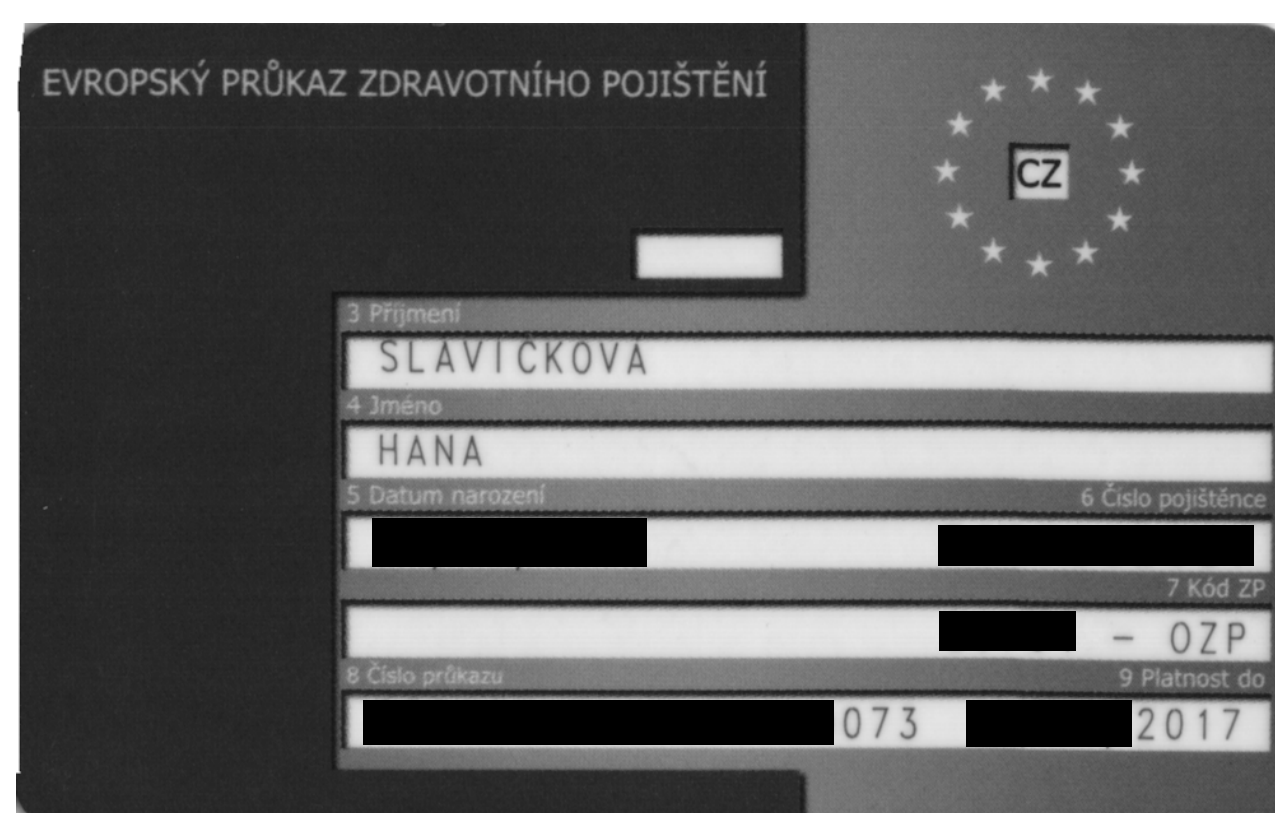
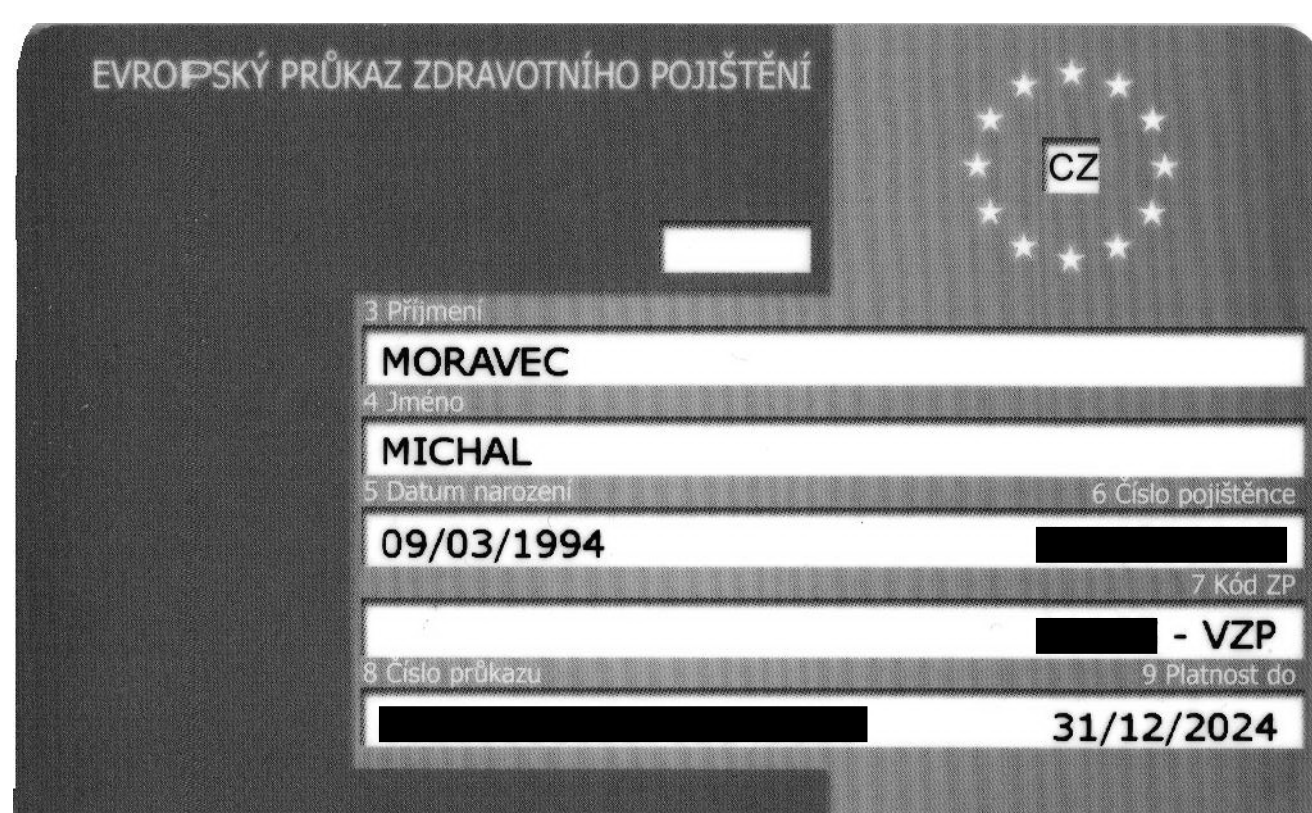
Hlavní funkce

- 1) Otočit kartu, aby výstupy nebyly vzhůru nohama nebo bokem
- 2) Zarovnat textová pole, které jsou posunutá nebo jinak poškozená vlivem špatného naskenování
- 3) Detekovat zadní stranu karty, na které se neprovádí detekce
- 4) Detekovat správný počet textových polí (standardně 5)
- 5) Výstupem jsou buď textová pole (5 výstupů) nebo jednotlivé informace na nich (7 výstupů)
- 6) Případně provést postprocessing, který zvýrazní kontrast mezi černým textem a bílým pozadím textového pole

Blokový diagram



Karta před filtrací



Textová pole po filtraci

Příjmení	MORAVEC
Jméno	MICHAL
Datum narození a číslo pojistnice	09/03/1994 [redacted]
Kód ZP	[redacted] - VZP
Číslo průkazu a platnost do	[redacted] 31/12/2024

Příjmení	SLAVICKOVA
Jméno	HANA
Datum narození a číslo pojistnice	[redacted] [redacted]
Kód ZP	[redacted] - OZP
Číslo průkazu a platnost do	[redacted] 073 [redacted] 2017

Příjmení	HORSKÝ
Jméno	JAROSLAV
Datum narození a číslo pojistnice	[redacted] [redacted]
Kód ZP	[redacted] - VZP
Číslo průkazu a platnost do	[redacted] 287 [redacted]