Portál pre odborné publikovanie ISSN 1338-0087

Spracovanie štruktúrovaných dát

Körösi Ladislav · Elektrotechnika, Informácie pre autorov

03.12.2014

V rámci riešenia projektu „Prenos, ukladanie a spracovanie štruktúrovaných“ boli vyvíjané softvérové moduly pre získavanie a prenos neštruktúrovaných dát z vizuálneho systému a tiež pre štrukturalizáciu a kategorizáciu získaných dát. Obrazové dáta boli získavané prostredníctvom vizuálneho systému Cognex In-sight Micro. V prvej fáze spracovania bolo potrebné aplikovať na získané dáta metodiku rozpoznávania obrazu – OCR (Optical character recognition). Takto získané textové dáta bolo možné následne kategorizovať na základe definovaných kľúčových slov. Počas vývoja funkčných modulov navrhnutej aplikácie boli aplikované metódy umelej inteligencie s použitím umelých neurónových sietí (UNS). Cieľom článku je stručne oboznámiť čitateľov s vytvorenými modulmi so zameraním sa na spracovanie údajov pomocou UNS.

Úvod

Cieľom navrhnutého riešenia bol komplexný proces získania, prenosu, spracovania a uloženia štruktúrovaných údajov. Na získanie nespracovaného 2D obrazu bola použitá kamera Cognex, ktorá v spojení s PLC ukladala obrazový materiál na PC. Pomocou nadstavbového softvéru bol obrazový materiál spracovaný pre získanie relevantných údajov. Špecifickým problémom spracovania obrazu je príprava dát pre dolovanie a reprezentáciu týchto údajov. Aj keď bod získaného 2D obrazu je numerického dátového typu, nie je vhodné dolovať celú obrazovú maticu ako celok. Obvykle môže byť pôvodný obraz skreslený alebo zašumený. Predspracovaním obrazu a extrahovaním vyššej úrovne informácií z obrazovej matice sa môže potlačiť šum a skreslenie. Okrem toho extrakcia vyššej úrovne informácií umožňuje porozumenie obsahu obrazu.

Obr. 1: Vizuálny systém Cognex In-sight Micro 1100

Vo všeobecnosti existujú nasledovné dátové typy:

- časové rady (zvuk, prevádzkové údaje, monitorovanie, lekárske údaje, …),
- obrazy (2D, 3D),
- video,
- text (rukopis, dokumenty),
- logy serverov a webové dokumenty.

Metódy vhodné pre vyššie vymenované typy dát:

- analýza časových radov,
- dolovanie obrazov,
- dolovanie videa,
- dolovanie textu,
- dolovanie Webu.

Pred experimentom dolovania dát je potrebné pripraviť dáta takým spôsobom, aby boli vhodné pre proces dolovania dát. Operácie pre prípravu dát možno rozdeliť nasledovne:

- spracovanie dát,
- normalizácia,
- spracovanie zašumených, neurčitých a nedôveryhodných informácií,
- spracovanie chýbajúcich dát,
- transformácia,
- kódovanie,
- abstrakcia.

Zohľadnením vyššie opísaných bodov boli navrhnuté funkčné moduly spracovania obrazového materiálu.

Opis modulov

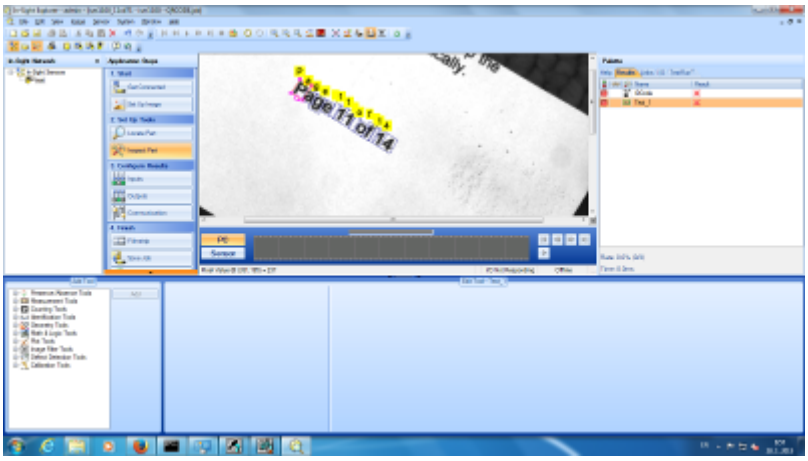
Grafické rozpoznanie textu sa skladá z dvoch hlavných častí:

- Predspracovanie – identifikácia jednotlivých písmen.

- Neurónová sieť – rozpoznanie znaku.

Identifikácia znaku sa skladá z viacerých krokov, ktoré budú opísané nižšie. V konečnom dôsledku, ale je potrebné získať rám okolo každého jedného znaku. Tento rám aj s jeho obsahom je možné potom použiť na vstup do UNS v procese učenia alebo vybavovania. Získaný obrázok z tlačených dokumentov nie je vhodný na rozpoznávanie textu s UNS vo forme získanou z kamery Cognex. Obrázok obsahuje rôzne nedokonalosti, šumy, ktoré treba odstrániť. Cieľom je získať z obrázku vstupné údaje vhodné pre UNS. Najprv sa segmentáciou rozdelí obrázok textu na jednotlivé znaky, z ktorých digitalizáciou sa získa vhodný vstupný vektor pre vstup UNS. Aby sa dal rozpoznať text pomocou UNS, vstupný obraz sa musí rozdeliť na menšie časti. Je potrebné nájsť riadky, potom v riadkoch slová a nakoniec rozdeliť slová na jednotlivé znaky. Predspracovanie údajov pozostáva z troch procesov:

- Predspracovanie obrázku
- Segmentácia obrázku
- Digitalizácia



Obr. 2: Prostredie Cognex In-sight explorer – trénovanie OCR

Modul „Segmentácia“

Prvým krokom predspracovania vstupných údajov je predspracovanie obrázku. Cieľom je zvýšiť kvalitu vstupných údajov a tým zvýšiť aj presnosť rozpoznania textu. Najčastejšie používanými metódami predspracovania obrazu pri získaných obrazových dokumentoch sú odstránenie šumu a binarizácia, ktoré boli implementované ako submoduly pre automatizáciu spracovania neštruktúrovaných údajov. Odstránenie šumu výrazne pomáha procesu OCR tým, že z pôvodného obrázku odstráni zbytočné elementy, ktoré nemajú vplyv na informáciu v dokumente.

Cieľom segmentácie je nájsť ostré hrany medzi znakmi, tak aby chybné vzory neboli zatriedené. Segmentácia môže zabezpečovať normalizovanie veľkosti a eliminovanie uhlov natočenia znakov. Segmentáciou je zabezpečené aj rozdelenie dokumentu na riadky a riadkov na slová. V rámci modulu bola riešená problematika segmentácie na riadky, segmentácie na slová ako aj segmentáciu na znaky.

Segmentácia na riadky

Úlohou bolo nájsť riadky v získanom dokumente. Na správne fungovanie algoritmu

musia byť riadky rovnobežné s horizontálnou osou, čiže natočenie nie je neprípustné alebo malo by byť minimálne. Na segmentáciu na riadky bol v prvom kroku použitý horizontálny histogram. Čím viac slov obsahuje riadok, tým je hodnota histogramu pre daný riadok väčšia. Nulové hodnoty histogramu medzi riadkami nemôžu byť jednoznačne považované za medzery, lebo kvôli šumom sa medzera niekedy ukazuje ako nenulový bod. Takým spôsobom by mohlo dôjsť aj k rozdeleniu písmen i alebo j a ich bodiek, alebo písmena a mäkčeňu, keď sú viac vzdialené od seba.

Segmentácia na slová

Úlohou algoritmu bolo rozdeliť riadky na slová. V tomto prípade sa použil vertikálny histogram. Princíp je podobný ako pri segmentácii na riadky. Za začiatok slova sa považovalo miesto, kde je hodnota histogramu väčšia ako stanovená konštanta. Slovo musí mať preddefinovanú dĺžku dl_slova obrazových bodov, a medzi dvoma slovami musí byť minimálne dl_medzera obrazových bodov.

Segmentácia na znaky

Algoritmus mal za úlohu rozdelenie slov na písmená. Postup je podobný ako v prípade rozdelenia riadkov na slová. Rozdielom je, že nie je nastavená hraničná hodnota na minimálnu šírku písmena a ani minimálna šírka medzi písmenami. V prípade veľkých písmen by segmentácia vyhovovala, ale v prípade malých písmen môže mať segmentácia rôzne výsledky. Negatívny vplyv na rozpoznávanie znakov majú hlavne rozličné výšky znakov. Na zabránenie chybnému rozpoznaniu sa použila normalizácia znakov.

Modul „Binarizácia“

Proces digitalizácie je veľmi dôležitý pre umelé neurónové siete používané v OCR. V tomto procese je na vstupný obrázok vzorky položené binárne okno (sieť reprezentujúca maticu), ktorá tvorí vstup do systému OCR. Cieľom bolo dostať tieto informácie do zmysluplnej formy pre UNS. Preto je pre každý čierny bod obrazu priradená hodnota +1 a pre každý biely bod 0 a tak vytvorená binárna matica. Digitalizácia obrázku do binárnej matice týmto postupom zaručuje nezávislosť rozpoznania od veľkosti znakov. Týmto spôsobom je zabezpečená jednotnosť rozmerov vstupných vzorov čo je postačujúce pre UNS. Binarizáciou boli rozdelené obrazové body vstupného obrazu do dvoch skupín: v jednej sú body znakov a v druhej body pozadia. V procese binarizácie sa použil jeden globálny prah na celý dokument.

Modul „Orezávanie“

Šum v krajných okrajoch získaného obrazu môže výrazne ovplyvniť výsledky segmentácie a aj výsledky rozpoznávania. Pri digitalizácii často vznikne na okrajoch dokumentu defekt (tzv. čierna čiara). Vzniká to pri kopírovaní a skenovaní dokumentu do digitálne formy. Submodul orezania mal za úlohu odstrániť zo vstupného dokumentu okrajové časti dokumentu. Obraz sa začne prehľadávať 2% od okraju vertikálne a 1% horizontálne. Algoritmus „orezáva“ stĺpce a riadky, a spočíta koľko bodov znaku sa nachádza v stĺpci a v riadku. Ak je tento počet menší ako pól percenta, tak sa odstráni aj tento stĺpec (riadok). Takto sa postupne odstraňujú ľavá, pravá, vrchná a dolná časť obrazu pokým nie je čo odstrániť.

Submodul UNS

Na rozpoznávanie znakov bola použitá viacvrstvová perceptrónová sieť (MLP – multi layer perceptron). Daná UNS bola zvolená na základe jej univerzálnemu použitiu, množstvu algoritmov učení, presnosti a ľahkej implementovateľnosti. Navrhnutá a implementovaná MLP sieť má tri vrstvy, ktoré postačujú na zakódovanie písmen. Počet neurónov vo vstupnej vrstve závisí od veľkosti vstupného vektora, ktorý je normovaný. Vstupom do neurónovej siete je binarizovaný znak, ktorý sa z maticovej formy (binárneho okna) pretransformuje do vektorovej podoby (stĺpcového vektora).

Ako už bolo zmienené program je univerzálny a preto je možné použiť binárne hodnoty alebo hodnoty z rozsahu 0 až 1. Neuróny v skrytej a výstupnej vrstve mali nastavené sigmoidálne aktivačné funkcie. Neuróny vo vstupnej vrstve boli lineárne. Počet skrytých neurónov bol stanovený na 2*n – 1, kde n je veľkosť vstupného vektora. Počet výstupných neurónov je pevne daná veľkosťou trénovanej množiny. V prípade učenia a rozpoznania kombinácií znakov A, B a C by UNS mala 3 výstupy. Pri trénovaní sa pre znak A nastaví vzor výstupu [1 0 0], pre znak B výstup [0 1 0] a pre znak C vektor [0 0 1]. Základným algoritmom z ktorého sa vychádzalo bol algoritmus spätného šírenia.

Obr.3: OCR pomocou UNS

Riešenie niektorých problémov pri rozpoznávaní

Na overenie kvality aplikácie boli vykonané rozsiahle testy. Tieto testy potvrdzujú dostačujúcu kvalitu rozpoznávania pomocou inteligentných metód rozpoznávania. Počas riešenia projektu boli riešené a odstránené aj nasledovné problémy pri rozpoznávaní:

Problém dvoj – objektových znakov

Algoritmus identifikoval prepojené objekty na obrázku. Problém nastal pri znakoch, ktoré sa skladajú z viacerých častí ako sú napr. písmena „i“ a „j“.

Spojené písmená

Niekedy pri použití veľmi malého písma v získavanom dokumente boli znaky tak blízko seba, že sa nemohli považovať za jeden objekt. Identifikácia takýchto objektov ako aj ich separácia bola riešená zadefinovaním dĺžky slova ako aj písmena.

Šum

Šum je najčastejší problém pri akomkoľvek spracovaní obrazu. Existuje na jeho odstránenie mnoho metód a filtrov. Vzhľadom na to, že sa spracovávali len s čierno – biele obrázky, tak bolo spracovanie jednoduchšie. Šum bol identifikovaný tak, že okolo seba mal už len biele body.

Záver

Navrhnutá metóda spracovania neštruktúrovaných dát a vytvorená aplikácia využívajúca umelé neuónové siete umožňuje kvalitné spracovanie získaného obrazu a identifikáciu relevantných údajov zo získaných dokumentov. Dosiahnutie vyššej kvality rozpoznávania zabezpečuje pridaná automatická slovníková kontrola s možnosťou manuálnej korekcie. Ďalšou výhodou vytvorenej aplikácie je integrácia možného doučenia použitých neurónových sietí na nové typy písma ako aj slov.

Poďakovanie

Tento článok vznikol vďaka podpore v rámci OP Výskum a vývoj pre projekt Prenos, ukladanie a spracovanie neštruktúrovaných dát, ITMS 26220220075, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Spoluautorom článku je Leo Mrafko.