Optikai karakterfelismerés

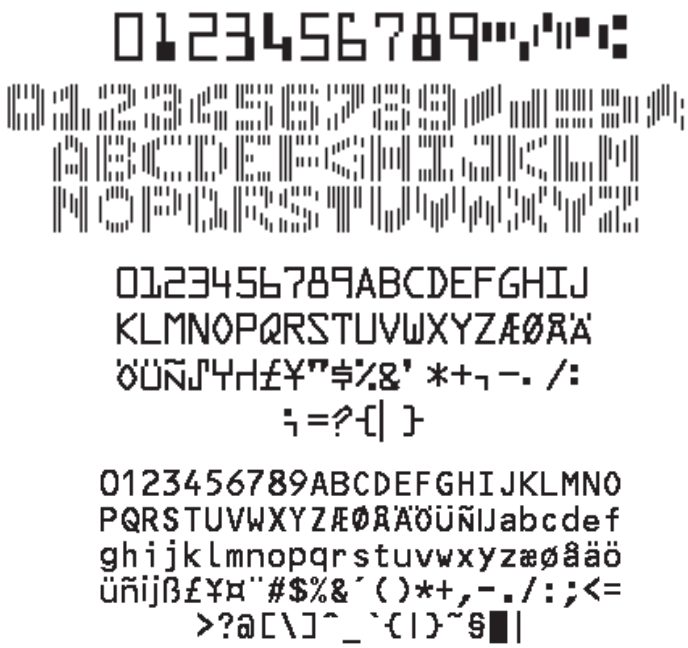
Az optikai karakterfelismerés feladata

A különböző formátumú dokumentumok kezelésének egyik speciális esete, amikor a kezelendő dokumentumok még nem állnak rendelkezésre elektronikus formában. Ebben az esetben szinte mindig arról van szó, hogy a dokumentumok kinyomtatva, papír alapú hordozón jelennek meg. Szövegbányászati tevékenység végzéséhez értelemszerűen digitalizálni kell a még nem digitalizált, papíron, nyomtatásban vagy írásban meglévő dokumentumokat, azaz a képként érzékelt dokumentumot szövegfájl formátumba kell átalakítani, hogy abban az után elektronikusan szerkeszthető és feldolgozható legyen. Ebben a szituációban kap szerepet az *optikai karakterfelismerés* (optical character recognition, OCR), amely így szövegbányászati előfeldolgozásnak tekinthető. Az optikai karakterfelismerés a mesterséges intelligencia jelfeldolgozó és generalizációs képességeit kiaknázva képes magas hatékonysággal nyomtatott, papír alapú dokumentumokon lévő karaktereket felismerni.

Az alap probléma itt az, hogy a nyomtatott papír alapú dokumentumok esetében nagy zajaránnyal kell megküzdeni annak érdekében, hogy a releváns információt kihámozzuk az érzékelt képi jelek és minták közül. Nyomtatott dokumentum esetében ilyen zajnak tekinthető például egy apró folt a papíron, tintaelmosódás, tintahiány, homályos háttér, apró gyűrődés a papíron, túl közeli vagy egybeolvadó betűk, betű dőlésszögének ingadozása stb. Kézírás esetén a kihívás még nagyobb, hiszen itt a személyiségjegyek sokszínűségéből adódó írásminták kavalkádjából kell kihámozni a karaktereket. Mind a nyomtatott, mind pedig a kézírásos esetben az optikai karakterfelismerő rendszer egy tanulási fázist követően képes olyan mintákat is osztályozni (értsd a megfelelő karaktert felismerni), amelyekkel a tanulási fázisban nem találkozott, tehát megvan a szükséges generalizációs képessége.

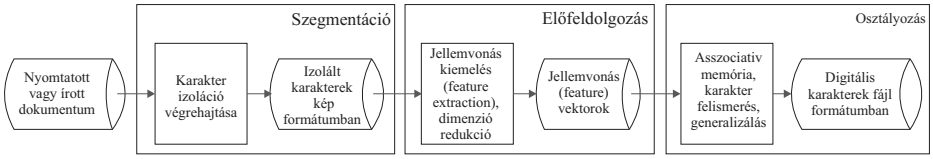
Az első üzleti alkalmazók a bankok voltak, ők használtak először optikai karakterfelismerő rendszereket. Kezdetben speciális karaktereket dolgoztak ki annak érdekében, hogy a karakterfelismerő rendszer dolgát megkönnyítsék. Így alakult ki az 1. ábrán látható többféle speciális betűtípus: E-13B (amerikai és kanadai bankok); CMC-7 (francia bankok); OCR-A, OCR-B.

Az OCR-A betűtípus az 1960-as években alakult ki, amikor már érett karakter felismerő rendszerek álltak rendelkezésre, de nem voltak kellően hatékonyak. Néhány karaktert jól láthatóan eltorzítottak, egyedivé tettek, így könnyítve meg a karakterfelismerő rendszerek dolgát. Az OCR-rendszerek alapvetően két részből



1. ábra. Néhány OCR-t támogató betűtípus, fentről lefelé E-13B, CMC-7, OCR-A, OCR-B

állnak: egyrészt a szkennelő fejből, amely a dokumentum egészét vagy részeit beszkenneli, másrészt pedig magából a mesterségesintelligencia-szoftverből, ami elvégzi a beérkezett minták osztályozását, azaz magát a karakterfelismerést. A nyomtatott latin vagy latin-rokon karakterek felismerése megoldott problémának tekinthető, az OCR-rendszerek igen hatékonyan képesek felismerni ezeket. A kézírás felismerése azonban még napjainkban is aktív kutatási terület, hiszen ez jóval összetettebb feladat. A 2. ábra az optikai karakterfelismerés folyamatát mutatja be.



2. ábra. Az optikai karakterfelismerés folyamata

Szegmentáció

A szegmentáció során a karakterek közötti éles határ megtalálása a cél annak érdekében, hogy téves minták ne kerüljenek osztályozásra (pl. két fél karakter). A szegmentáció feladata lehet az is, hogy a karakter-dőlésszögeket, karaktermérete-ket normalizálja. Sok esetben a szöveges dokumentumokban nem csak karakterek vannak, hanem képek és egyéb, a felismerés szempontjából nem lényeges szimbólumok. A szegmentáció további feladata tehát az is, hogy az ilyen, számunkra nem releváns grafikus objektumok közül kiszűrje a csak karaktereket tartalmazó szöveges részeket. A 3. ábra a szegmentáció egy esetét szemlélteti.

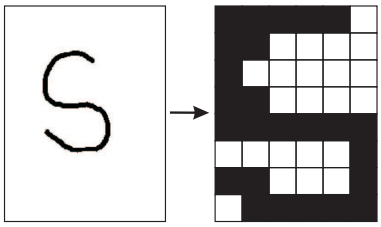
3. ábra. Példa szegmentációra

Optikai előfeldolgozás

Az előfeldolgozás a bemeneti minta komplexitásának csökkentésére szolgál, és annak legjellemzőbb vonásait elemi ki. Különösen nagy jelentősége van a kézírás felismerésekor, ugyanis az írott betűk jóval komplexebb mintákat alkothatnak, mint a nyomtatott betűk. A jellemzőkiemelés során a komplexitás úgy csökken, hogy közben a legjellemzőbb információk megmaradnak és ezáltal a későbbi feldolgozás számításigényét redukálhatjuk. Ez a folyamat tulajdonképpen egy komplexitáscsökkentéssel járó digitalizáció. A 4. ábra egy egyszerű digitalizálási módszert mutat, amikor az analóg jelre egy mátrixot reprezentáló rácshálót illesztünk, és amelyik cellán átmegy az analóg karakter, az az elem a mátrixban 1 értéket vesz fel (fekete), egyéb esetben pedig 0-t (fehér).

Osztályozás

Az osztályozás során történik meg a tényleges karakterfelismerés. A karakterfelismerő módszer a bemeneti jellemzővektor alapján dönti el, hogy az ismert karakterek közül melyikre hasonlít a legjobban a bemeneti vektor. Így a karakterfelismerési probléma egy asszociatív memóriát igénylő feladat, amelynek során a tárolt memóriaelemek közül kell előhívni azt, amely a bemeneti mintának legjobban megfelel. Az asszociatív memória elsősorban a mesterséges neurális hálózatok-



4. ábra. Egyszerű digitalizálási módszer

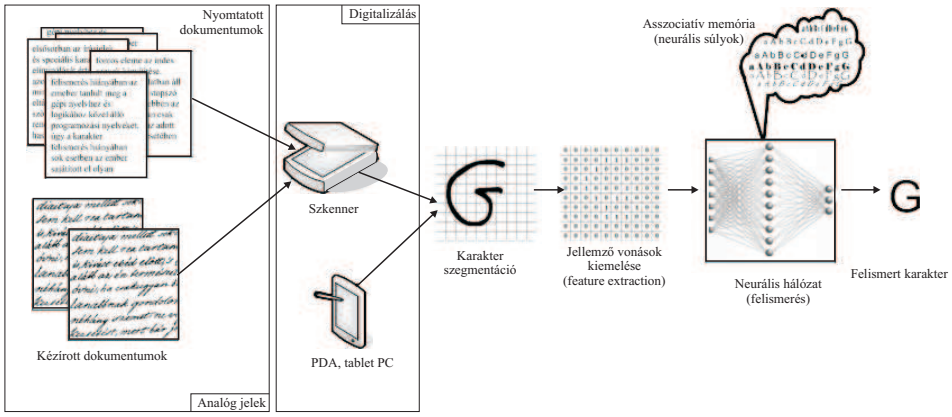
nál előforduló fogalom, ugyanis ilyen neurális hálózatokkal (pl. Hopfield-hálózat) nagy hatékonyságú asszociatívmemória-megoldások implementálhatóak.

Minthogy a karakterfelismerés alapvetően osztályozási probléma, ezért az 5. fejezetben bemutatott módszerek a karakterfelismerésben is alkalmazhatóak. Ettől függetlenül elmondható, hogy a neurális hálózat alapú karakterfelismerő megoldások a legelterjedtebbek az 1980-as évek óta, jelenleg azonban a szupportvektor-gépekkel (ld. a könyv 127. oldalán) történő karakterfelismerési és osztályozási problémák kutatása a legnépszerűbb.

Napjainkban nem csak a nyomtatott szövegek felismerése a feladat, egyre nagyobb igény mutatkozik az emberi kézírás felismerésére is. Ennek megvalósítása komolyabb apparátust igényel, mivel az írott karakterek felismerési problémájának komplexitása meghaladja a nyomtatott betűkét. A tablet PC-k és egyéb, az emberi kézírás befogadására képes eszközök terjedésével még flexibilisebb kapcsolat valósulhat meg az ember és gép közötti, ha sikerül kellően hatékonnyá tenni a kézírás-felismerő módzsereket. Egy IBM által elvégzett felmérésből kiderült, hogy a 97% alatti hatékonyságú kézírásfelismerő-rendszereket a megkérdezett felhasználók használhatatlannak minősítették.

A természetes nyelvi feldolgozást motiváló tényező a karakter felismerés terén is erőteljesen érezteti a hatását, mely szerint nem az emberek kívánják megtanulni a gépek nyelvét, hanem igyekszünk olyan gépeket fejleszteni, amelyek képesek az emberi nyelvet, illetve kézírást felismerni. Ahogyan a természetes nyelvi felismerés hiányában az ember tanult meg a gépi nyelvhez és logikához közel álló programozási nyelveket, úgy a karakterfelismerés hiányában sok esetben az ember sajátított el olyan speciális írásmódot, amit a számítógép képes volt megérteni — lásd például a Palm személyi asszisztensekben (palmtop) elterjedt Graphiti karakterkészletet, amelyet a gép képes felismerni, de itt minden karaktert csak egy vonás lehet (Unistroke). Ez tehermentesítette a felismerőszoftvert a karakterszegmentáció problémájától.

A trend tehát a természetes emberi kézírás számítógépes felismerése felé is mutat a nyomtatott karakterfelismerés mellett (Faaborg, 2002). Az ember képes tolerálni az érzékelt karakterek felismerésekor azok színét, stílusát, dőlésszögét, pozícióját, méretét, méretarányát és az elmosódáoskból és egyéb vizuális torzulásokból adódó zajt. Olyan emberek kézírását is képesek vagyunk elolvasni, akikét még soha nem láttuk, akiknek egyedi stílusával még soha nem találkoztunk. Az ember elképesztő sebességgel képes az új karaktertípusokat stílusokat magáévá tenni és felismerni. Hatékony gépi intelligencia kifejlesztésekor ezeket a készségeket a számítógépnek is meg kell tudnunk tanítani, legalábbis egy elfogadható szintre fejleszteni. Mivel nincsenek olyan egzakt szabályok, amelyek algoritmust adhatnának egy kézírásos szöveg karaktereinek felismerésére, így jobbhíján a felügyelt tanulás jöhet szóba, amelynek az egyes karakterek felismerésekor tolerálnia kell az esetleges zajokat. Az 5. ábra a karakterfelismerés neurális hálóval történő felismerésének folyamatábráját szemlélteti. A felismerési folyamat a neurális hálózat példákon keresztüli betanítása után következhet. A neurális hálózat az úgynevezett jellemzők kinyerése révén képes végrehajtani a halmazszeparációs problémát, ami végül jó esetben a karakterfelismerésére vezet (Araokar, 2005).



5. ábra. Karakterfelismerés neurális hálózattal

Forrás:

S. Araokar. Visual Character Recognition using Artificial Neural Networks. *ArXiv Computer Science e-prints*, (cs/0505016), 2005.

A. J. Faaborg. Using neural networks to create an adaptive character recognition on system, 2002.