Multi-modális ember-gép kapcsolatok

Fazekas Attila, Szeghalmy Szilvia, Bertók Kornél, Sajó Levente

Debreceni Egyetem, Debreceni Képfeldolgozó Csoport, 4010 Debrecen, Pf.:12.
{attila.fazekas, szeghalmy.szilvia, bertok.kornel, sajo.levente}@inf.unideb.hu

Abstract. Az információs rendszerek használatának egy új módját jelenti a multi-modális ember-gép kommunikáción alapuló technikák használata. Ebben a cikkben egy rövid áttekintést kívánunk adni az ilyen technikán alapuló kommunikáció egy – jelenleg futó projektünk keretében – lehetséges megvalósításával kapcsolatban.

Keywords: Multi-modális ember-gép kommunikáció

1 Bevezetés

Az információs társadalom egyik alapvető igénye az információkhoz való hatékony hozzáférhetőség biztosítása. Az elmúlt évtizedben a technológia fejlődése egyre hatékonyabb eszközöket adott a kezünkbe az információ tárolására, rendszerezésére és lekérdezésére. Továbbá az élet számos területén megjelentek olyan eszközök, amelyek feladata a tárolt információk lekérésének biztosítása. Ezen eszközöket gyűjtőnéven információs rendszereknek nevezzük. Ebbe a kategóriába tartoznak az általános célú számítógépek, de azok a speciális számítógépek is, amelyek valamilyen jól definiált célra készültek, mint például a menetrendekkel kapcsolatos információk tárolása, visszakeresése és az esetleges foglalások, illetve jegyek on-line vásárlásának lebonyolítására.

Annak ellenére, hogy az információs rendszerek életünk számos területén jelen vannak a használatukkal szemben egyfajta ellenállás figyelhető meg. Ez egyrészt a technológia használatához szükséges ismeretek hiányának, másrészt az egyes rendszerek használati módjának különbözőségéből eredhet. Ez utóbbi azt jelenti, hogy életünk minden pillanatában újabb és újabb információs rendszer használatához szükséges ismereteket kell elsajátítanunk. Ráadásul az információs rendszerek egy-egy új generációjának megjelenése egyre rövidebb idő alatt következik be, és életünk egyre több területén hódítanak teret maguknak.

Az információs rendszerek használatának módja egyfajta kommunikációs nyelvnek a használatához hasonlít. Ha sok, különböző információs rendszerrel való kommunikációhoz szükséges nyelvet ismerünk, akkor az egy-egy újabb ilyen kommunikációs nyelv elsajátításához szükséges idő egyre kevesebb lesz. Továbbá, ha egy információs rendszer kommunikációs nyelvét napi szinten használjuk, akkor

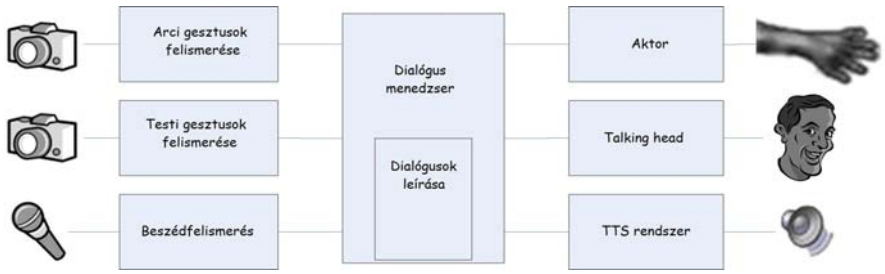
általában könnyebb lesz elsajátítanunk egy újabb generációjának a kommunikációs nyelvét.

A fentiek alapján teljesen nyilvánvaló, hogy jelenleg minden esetben a felhasználónak kell megtanulnia az adott információs rendszer kommunikációs nyelvét, azaz azt, hogy milyen formában adhatunk utasításokat a rendszernek, illetve milyen formában kapjuk meg a rendszerben tárolt információkat. A multi-modális ember-gép kommunikációval kapcsolatos kutatások alapgondolata az, hogy találjuk meg annak a módját, hogy a jövő információs rendszerei képesek legyenek a felhasználóval a számukra legtermészetesebb módon kommunikálni, azaz lehetővé tenni az emberi nyelv használatát. Egy felhasználónak nem kell újabb és újabb kommunikációs nyelvet elsajátítania, hanem elegendő „szóba elegyednie" az adott rendszerrel.

Minden szempont alapján ideális multi-modális ember-gép kommunikációs rendszer még nem készült el. Ezzel is magyarázható, hogy az ipari fejlesztések területén a kivárás a jellemző. A jelenleg futó projektünk keretében az ember-gép kommunikációban használatos vizuális jegyek digitális képfeldolgozási eszközökkel történő felismerésének problémájával foglalkozunk. Ennek következtében cikkünk fő célkitűzése az, hogy röviden vázoljuk egy általános rendszer felépítését és az egyes komponensek funkcionális szerepét. A projektben elért, a teljes rendszer szempontjából egy-egy komponensnek tekinthető részfeladatok megvalósításával kapcsolatos eredményeket külön cikkek keretében kívánjuk bemutatni.

2 Multi-modális ember-gép kommunikációs rendszer felépítése

Egy általános rendszer koncepciója az, hogy egy virtuális személyt valósítsunk meg, aki a kommunikáció szempontjából a lehető legtöbb tekintetben emberként viselkedik. Egy ilyen kommunikációt megvalósító általános rendszer felépítését az 1. ábra szemlélteti.



1. ábra. Egy multi-modális ember-gép kapcsolatot megvalósító általános rendszer.

A virtuális rendszert funkcionálisan két fő részre bonthatjuk: a kommunikáció lebonyolítását biztosító modulra, valamint az ember-gép kommunikációt szolgáló felület modulra.

A kommunikációért felelős modul szintén több kisebb komponensből épül fel. Mivel a rendszer feladatából adódóan egy többé-kevésbé ismert kommunikációs helyzetbe kerül, ezért erre a szituációra vonatkozó általános ismeretek strukturális tárolása elősegíti a rendszer emberszerűbb viselkedését. Ezek az ismeretek egyrészt az elképzelhető dialógusok reprezentálását, másrészt a kommunikáció nyelvének strukturális és szókincsére vonatkozó információk tárolását jelenti.

Az ember-gép kommunikációt megvalósító komponens több emberi kommunikációs csatornát felhasználva, multi-modális kapcsolatot tesz lehetővé. A bemeneti interfészek hardver elemei: webkamera, amely az emberi partner arcát, testét figyeli; mikrofon a partner hangjának rögzítésére. A webkamera interfészt képfeldolgozási módszereket megvalósító szoftver egészíti ki, amely képes felismerni a kommunikációban részt vevő ember arcán és testén megjelenő vizuális jeleket (neme, életkora, arci érzelmek, kézjelek). A beszédet a beszédfelismerő szoftver dolgozza fel, amely képes a kommunikáció során elhangzott kulcsszavakat detektálni, és a beszélő érzelmi „hangsúlyait” felismerni.

Az ember-gép kommunikáció kimeneti interfészének hardver egysége a hangszóró mellett a monitor, a szoftver komponense pedig a szövegfelolvasó szoftver mellett az érzelmek kifejezésére is képes, animált beszélő fej. [3]

2.1 Arci jellemzők felismerése

Az emberi arc önmagában is egy információhalmaz, amelyből mi, emberek bármikor ki tudjuk nyerni az életkort, nemet, érzelmi állapotot, stb. Ahhoz azonban, hogy mindezt fel tudjuk használni a számítógéppel történő kommunikációban, a felismerést számítógéppel kell végeznünk, ami összetett képfeldolgozási feladat. A fent említett információk (érzelem, nem, életkor, kulturális közeg, tekintet iránya) kinyerésére statisztikai tanuló algoritmusokat alkalmazunk (Support Vector Machine). A gyakorlatban azonban számos előfeldolgozási lépést kell tennünk ahhoz, hogy az arcokból információt nyerhessünk ki: meg kell találnunk az ember arcát a képen (arcdetektálás), valamint követnünk kell azt mozgás során (arckövetés), ráadásul mindezt, beleértve a tanuló algoritmusok osztályozását is, valós időben kell végeznünk. A rendszer a felismerést jelenlegi állapotában másodpercenként 2-3 alkalommal tudja végrehajtani. A rendszer ezen komponenseinek részletei megismerhetők a [2]-ben. Mivel az érzelmek nem váltakoznak gyorsan, a videófolyamot felhasználva a korábbi detektált érzelmek alapján átmeneti valószínűségeket figyelembe véve a módszer még robosztusabbá tehető.

A teljesség igénye nélkül most csak az emberi fél tekintetének meghatározását ismertetjük vázlatosan. Ebben az esetben a pupilla középpontját kell megtalálnunk. Ehhez első lépésben a szivárványhártyát keressük meg egy előre betanított Viola-

Jones detektor segítségével. Majd a detektor által visszaadott szivárványhártya képeken a pupilla egy egyszerű küszöbölés segítségével egyértelműen lokalizálhatóvá válik. A szivárványhártya detektor segítségével a pislogás is modellezhető, mivel a detektor az esetek túlnyomó többségében csak akkor nem talál szivárványhártyát, ha a felhasználó szeme éppen csukva van.

2.2 Testi jellemzők felismerése

A kommunikáció során a mozdulatoknak, testtartásnak is fontos szerep jut. Bár ezek a gesztusok gyakran nem is tudatosak, a kutatások rávilágítottak arra, hogy a partner reagál ezekre a gesztusokra is [1]. A testi jellemzők közül jelenleg a rendszer a fej állását tudja viszonylag robusztus módon meghatározni. A kézmozdulatok követése és a gesztusfelismerés viszont messze túlmutat egy egyszerű objektumkövetésen. A kezek változó megjelenési formáival, a gyakori átfedésekkel, takarásokkal, de még az öltözködésben való eltérésekkel is meg kellene a rendszernek küzdenie. A jelenlegi módszerekkel azonban csak erősen megkötött szituációkban lehet elfogadható eredményeket elérni.

A fej térbeli helyzetének meghatározására szolgáló eljárásunk a következő módon működik: A detektált arci jellemzők síkbeli koordinátái alapján a POSIT eljárás segítségével módosítjuk egy előre rögzített térbeli fejmodell alappontjait. Az eljárás eredményeként megkapjuk az emberi fej térbeli modelljét leíró forgatások szögeit, illetve eltolások vektorait.

2.3 Beszéd felismerése

A jelenleg az irodalomban ismert beszéd felismerő rendszerek alapjában véve bizonyos kulcsszavak felismerését teszik lehetővé, ezért egy ember-ember kommunikációt megközelítő párbeszéd kialakulásához mindenképpen szükség van arra, hogy a kommunikációs szituációval kapcsolatos információk a lehető legteljesebb formában rendelkezésre álljanak. A lehetséges dialógusok szerkezetének ismerete lehetővé teszi, hogy tetszőleges szöveg felismerése helyett, elegendő legyen egyes izolált kifejezéseket azonosítani, amelyek a kommunikációban szereppel bírnak.

2.4 Beszélő fej

A virtuális ember komponens kimeneti interfésze a beszélő fej, amely az emberrel szemben lévő képernyőn jelenik meg. Két kimeneti csatornát használunk: a szintetizált beszédet, valamint a beszélő fejen megjelenő gesztusokat. A beszédszintetizálás a ProfiVox TTS rendszerrel történik [4], míg a megjelenő animált fej a CharToon rendszerre épült [5]. A beszédszintetizátor minden verbális megnyilvánulás esetén előállít egy hanghullám állományt, amely a szintetizált beszédet tartalmazza, valamint egy időparaméterekkel ellátott fonéma szekvenciát,

amely azt az információt tartalmazza, hogy melyik pillanatban milyen fonéma hallható. A beszélő fej szájának animálásához definiáltuk a magyar nyelv egyes fonémáinak kiejtésekor megjelenő arcokat, ezeket a fonémához tartozó vizémának nevezzük. A beszéd animálása tehát a képernyőn megjelenő fejhez definiált vizémák lináris interpolálása olyan módon, hogy a hangállomány párhuzamos lejátszásakor a fonéma és vizéma párok a megfelelő pillanatban jelenjenek meg.

A másik csatorna a beszélő fej esetén az érzelmek kifejezése. Jelenleg 4 érzelmi

2. ábra. A beszélő fej érzelmi állapotai: természetes, vidám, szomorú, unott

állapotot tudunk megjeleníteni: természetes, szomorú, vidám, unott (2. ábra). Mindemellett a fej véletlenszerűen kisebb arci gesztusokat tesz (pislogás, szájhúzogatás), ezzel is élethűbbé téve a fej viselkedését. A véletlenszerű folyamatok vezérlésére a fej esetén a Perlin-zaj függvényt használjuk, mivel az más zajfüggvényeknél jobban közelíti a valós világban előforduló természetes viselkedésmintákat.

3 A rendszer működése

Egy, ebben a cikkben vázolt multi-modális ember-gép kommunikációs rendszer működése a következőképpen képzelhető el. A kamerák egy része a kommunikációban résztvevő ember arcát, míg a másik részük a felső testét „figyeli”. Az arcon látható és a karok és kezek által kifejezet gesztusok felismerésének eredménye kombinálódik a beszéd felismerő által szolgáltatott érzelmi jellemzőkkel. A beszédfelismerő által felismert kulcsszavak alapján a dialógus menedzser a lehetséges kommunikációs szituációk ismeretében rekonstruálja az elhangzottakat, aminek jelentését módosíthatják a gesztus felismerőből származó információk. A dialógus menedzser meghatározza a gép reakcióját, amit a beszélő fej az arcán érzelmeket kifejezve, az érzelmi állapotának megfelelő hangon a beszédszintetizátor el is mond.

4 Összefoglalás

A jelenleg is futó projektünk keretében az arci- és testi gesztusok felismerésének lehetőségét kívánjuk megvizsgálni. Arra való tekintettel, hogy ez a kutató/fejlesztő munka még folyamatban van, ezért csak a projekt jelenlegi állapotát tükröző eddigi eredményeket szeretnénk bemutatni a konferencián további cikkek keretében. Ezen cikk szerepe elsősorban a közös fogalmi háló és rendszer általános koncepcióját volt hivatott ismertetni.

References

1 B. Budai, A közvetlen emberi kommunikáció szabályszerűségei, Budapest, Animula, 2001.
2 A. Fazekas, Face analysis and the simultaneous processing of audio and video channels in human-computer interaction, in Proc. of NAES-FINSSE 2010, 9-13 June, 2010, Oulu, Finnland, 18-19.
3 Gy. Kovács, Zs. Ruttkay and A. Fazekas, Virtual Chess Player with Emotions, in Proc. of Fourth Hungarian Conference on Computer Graphics and Geometry, from 2007-11-13 to 2007-11-14, Budapest, Hungary, 182-188.
4 G. Olaszy., G. Németh, P. Olaszi, G. Kiss, G. Gordos, "PROFIVOX - A Hungarian Professional TTS System for Telecommunications Applications", /International Journal of Speech Technology/, Volume 3, Numbers 3/4, December 2000, 201-216
5 Zs. Ruttkay, H. Noot. Animated CharToon Faces. Proc. of NPAR 2000 - First ternational symposium on Non Photorealistic Animation and Rendering, June 2000, 91-100