VIZUÁLIS FIGYELEM MODELLEZÉSE

Tézisfüzet a Ph.D. disszertációhoz

Lázár Anna Kinga

Témavezető:
Dr. Roska Tamás
a Magyar Tudományos Akadémia rendes tagja

Konzulens:
Dr. Vidnyánszky Zoltán
a Magyar Tudományos Akadémia doktora

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
Multidiszciplináris Műszaki Tudományok Doktori Iskola

Budapest 2008

1.Bevezetés

Az egészséges ember figyelmi mechanizmusa oly természetességgel és könnyedséggel működik, hogy többnyire tudatában sem vagyunk annak, milyen összetett feladatról is van tulajdonképpen szó. Látásunk segítségével szinte világrajövetelünk pillanatától fogva informálódunk a külvilág objektumairól és eseményeiről, és - szerencsés esetben - életünk végéig egyik legfontosabb érzékszervünk marad. Fontossága és *trivialitása* miatt – hogy mennyire természetesnek éljük meg napról napra a látást, mint élményt – régóta és sokan foglalkoznak mind a megértésével, mind az „utánzásával", modellezésével. Habár eközben tagadhatatlanul rengeteg tudás és információ halmozódott fel, a teljes megértéstől még igen távol vagyunk: hogyan lesz a fotonok sejtingerléséből látási élmény, vizuális tudat.

Élményünk szerint aktuális környezetünk egészét folyamatosan észleljük, befogadjuk: tudjuk, hogy mi van fölöttünk, alattunk, tőlünk balra vagy jobbra, vagyis minden kis részletről folyamatosan informálva vagyunk látásunk révén. Ez – habár nem teljesen alaptalan -, elsősorban mégiscsak egy hamis érzés. Kísérletek tanulsága szerint környezetünk akár igen jelentős átalakulásokon is áteshet (pl. a falak piros színűről kék színűre változhatnak, vagy egy tárgy pont az orrunk előtt elhalványodhat és eltűnhet); ha a változás nem elég gyors, akkor az egészből mit sem veszünk észre. Ennek magyarázata pontosan abban rejlik, hogy valójában csak egy eltárolt

„képünk, reprezentációnk” van a minket körülvevő világról: ennek azonban több köze van az emlékezethez, esetleg eltárolt tudásunkhoz, mint a konkrét látáshoz. És mindennek természetesen igen jó oka van: ha minden pillanatban *valóban* feldolgoznánk a környezetünkből jövő összes információt, agyunk borzasztóan túl lenne terhelve *fölöslegesen*: a félig előttünk lévő könyvespolc alakja, mérete, a rajta lévő könyvek mindegyikének pontos címe, formája és színe, a szemközti ház tetején a cserepek, a szomszéd szobából látszódó falikép minden egyes részlete: mindez *egyszerre* biztosan nem fontos számunkra. Sőt, ha kicsit belegondolunk, valójában többnyire környezetünknek csak igen kis része fontos adott helyen, adott időben: a többi redundáns (a falon lévő tapéta ismétlődő mintája) vagy épp érdektelen (a szemközti ház tetőcserepei). Így tehát az „eltárolt reprezentáció módszer” kifejezetten előnyös. Ha meg az eltároltakhoz képest valami változás történik (pl. felgyullad a lámpa a szomszéd szobában, valaki bejön, stb.), akkor arról a megfelelő (úgynevezett „bottom-up”) figyelmi mechanizmus révén úgyis értesülünk.

És itt eljutottunk a vizuális figyelem fogalmához: ehhez a rendkívül fontos képességhez, amely valós időben, a teljes információhalmaz feldolgozása nélkül lehetővé teszi annak a kis vizuális részletnek a kiválasztását, megtalálását, amely adott időpontban éppen fontos a (mesterséges vagy élő) rendszer számára.

Ezen képesség evolúciós jelentőségét nehéz túlhangsúlyozni: gondoljunk csak a hirtelen megjelenő ragadozókra, az érett gyümölcs

megtalálására a sűrű bozótosban vagy a fán, vagy éppen a fajtársak jeleire.

Mérnöki szempontból tehát egy rendszer, amely '*figyel*', képes *valós időben* megtalálni a vizuális tér azon kis részletét, amely az adott rendszer számára *akkor és ott éppen fontos*. Ezzel igen jelentős számítási kapacitás spórolható meg - hiszen nem kerül feldolgozásra a redundáns/érdektelen adattömeg – ami által egyrészt a feldolgozás *minősége* javítható, míg a processzáláshoz szükséges *idő* csökkenthető. És persze az, hogy mi az aktuálisan *fontos*, rendkívül összetetté és bonyolulttá teszi a problémát.

Iskolánkban elsősorban neuromorf modellezés folyik, vagyis modelljeink alapjait az élő rendszerektől lessük el. Az ember figyelmi mechanizmusának alapvetően két összetevője van: egy önkéntelen, reflexszerű (úgynevezett „bottom-up"), ami például egy pirosan villogó lámpára irányítja figyelmünket az utcán, és egy tudatosan irányított (ún. „top-down"), aminek segítségével egy tömött fiókban megtaláljuk a kulcsunkat.

Doktoranduszi éveim alatt ezen bottom-up figyelmi rendszer megértése és modellezése volt a legfőbb feladatom. Ennek során elkészült egy neuromorf alapelveken nyugvó modell, amelynek szabad paramétereit emberi szemmozgás mérések alapján állítottam be. Az így elkészült modell 'jóságát', minőségét szintén emberi szemmozgásokhoz viszonyítva teszteltem.

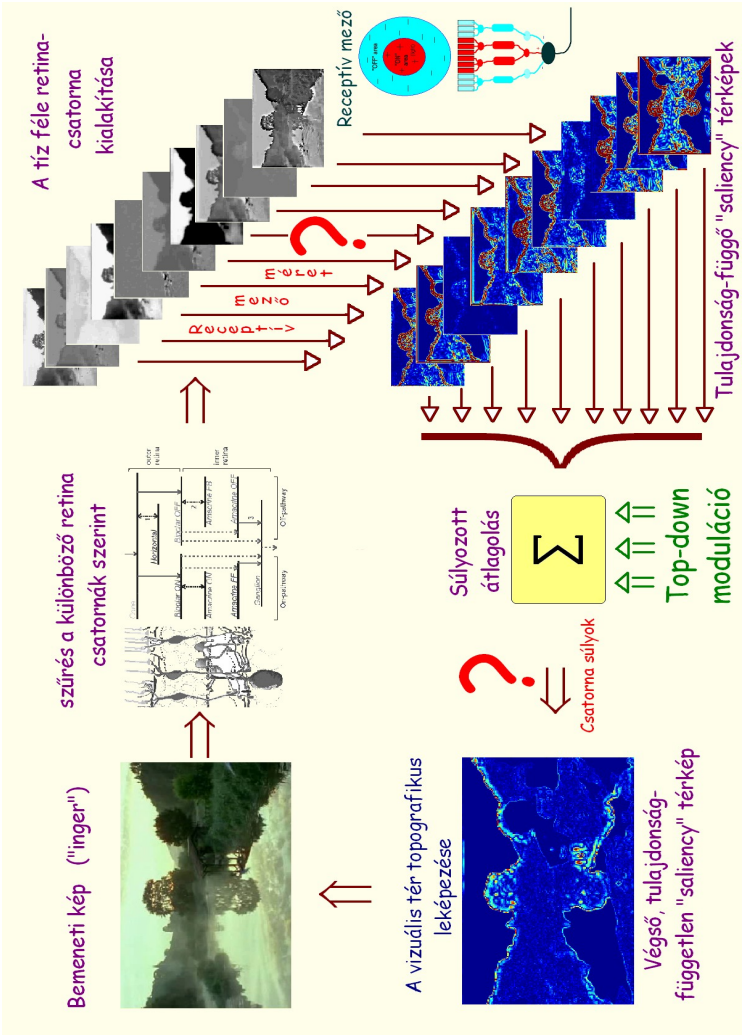
2.Vizsgálati módszerek

Kutatási témám különféle tudományágak együttes alkalmazását kívánja meg. Így első lépésként neurobiológiai tanulmányok során a látás alapjaival és a vizuális figyelmet alakító mechanizmusokkal ismerkedtem meg.

A modellkészítés lényege abban áll, hogy egy komplex rendszerből (pl. egy állat látórendszere) jól válasszuk ki a legfontosabb elemeket, azokat, amelyek a vizsgált rendszer általunk fontosnak tartott tulajdonságait alakítják. Ezért, ha ezek az elemek több rendszerben is azonosak, akkor kismértékű 'átjárás' lehetséges a modellek alapjait illetően. Így munkám alapjának általános értelemben a gerincesek látórendszere tekinthető. (A jelenlegi figyelmi *modellek* még távolról sem annyira pontosak, hogy pl. a majmok és az emberek közötti különbségek megjelennének bennük.)

A modell főbb lépései az 1. ábrán láthatóak. A bal felső sarokban látható bemeneti kép először is a különböző retinacsatornák szerint felbomlik (jobb felső kép). Ezután minden egyes csatorna kialakítja a saját ún. „saliency" (azaz „feltűnőségi") térképét[♠], amelyek a vizuális térnek az agyban keletkező topografikus leképezései (jobb alsó ábra). Ezen 'térképek' súlyozott átlaga fogja kialakítani a végső (ún. „final" vagy „master") *feltűnőségi térképet* (bal alsó ábra),

[♠] A „saliency" egy angol szó, jelentése: kiugró, feltűnő. A „saliency map" kifejezés a figyelemmel foglalkozó szakirodalomban igen elterjedt, bevett magyar fordítása azonban nincs. Én a dolgozatban ezt a kifejezést „feltűnőségi térképnek" vagy „feltűnőségi leképezésnek" hívom.



1. ábra: A bottom-up figyelmi mechanizmus működése. Első lépésként a bejövő kép felbomlik az egyes retina csatornák szerint, kialakítva ezzel a bejövő kép tíz topografikus leképezését (jobb felső sarok). Ezután mindegyik csatorna létrehozza a saját 'feltűnőségi' leképezését (jobb alsó rész), amelyek súlyozott összege fogja adni a végső, „master" térképet (bal alsó ábra), amelynek legaktívabb pontja vonzza a figyelmet.

amely a vizuális térnek szintén egy topografikus kódolása, és azt mutatja meg, hogy a fizikai világ megfelelő pontjai mennyire feltűnőek, kiugróak, mennyire ütnek el a környezetüktől. Ezen térkép legnagyobb értékű pontja vonzza leginkább a figyelmet, a fizikai világ neki megfelelő részlete fog az éleslátás középpontjába kerülni.
A fenti modellt (Borland) C++ nyelven valósítottam meg.
Az első főbb lépés a retinacsatornák vizsgálata és elkészítése volt, amelyben Bálya Dávid segített sokat. A modell egy CNN (Cellular Neural/Nonlinear Network) szimulátoron fut, amely szimulátort szintén Borland C-ben írtam meg. Az egyes retinacsatornákat meghatározó, (diffúziót, időbeli lefutást, stb.) kialakító konkrét paramétereket szintén Bálya Dávidtól vettem át. A modell további részei – természetesen témavezetőm és konzulensem útmutatásai alapján – saját munka eredményei.
A retinamodell alapelve röviden a következő: a retina minden sejtrétegének (foto-receptorok, horizontális-, bipoláris-, amacrine- és ganglion sejtek rétege) egy-egy CNN-réteg felel meg. Az egyes sejtrétegek tulajdonságai (dendritfák átlagos átmérője, a sejtválaszok időbeli lefutásai, stb.) megfelelő CNN template-ekkel és paraméterekkel közelíthetőek. Az így beállított CNN rétegek közötti kapcsolatok (serkentés/gátlás, időbeli késleltetés, diffúziós állandók, stb.) szintén úgy vannak meghatározva, hogy minden CNN-réteg minél jobban közelítse a neki megfelelő retinális réteg válaszát. A retina *időbeli* tulajdonságait egy súlyozott, cirkuláris memória segítségével modelleztük: az újonnan feldogozott frame mindig a legrégebbit írta felül, és a retinacsatorna outputja a cirkuláris

memória teljes tartalmának súlyozott, pixelszintű összegzése révén jött létre.

Következő lépésként az egyes retinacsatornákra „ülő” feltűnőségi térképeket kell létrehozni. A különböző csatornákhoz tartozó leképezéseket különböző méretű receptív mezők (RF) alakítják ki. A receptív mezők a vizuális feldolgozás elején kör alakúak, majd egyre feljebb haladva az agyi hierarchiában, alakilag egyre összetettebbekké, méretileg pedig egyre nagyobbakká válnak. Mivel kutatásaim elején az egyes csatornákhoz tartozó *receptív mező méretek* (pontosabban a mező-méretek eloszlásai) ismeretlenek voltak, ezt a méretet az alapmodellben egyszerűen billentyűzetről beállítható szabad paraméterként hagytam. (Az ábrán az ismeretlen paramétereket piros kérdőjelek mutatják.)

A másik igen fontos, de szintén ismeretlen paraméter a csatornafüggő feltűnőségi térképek *súlyozása*, amely súlyozás mellett a végső, azaz „master” leképezés kialakul (ábra alsó része).

Ezeket a paramétereket embereken végzett szemmozgás mérések alapján határoztam meg. Ehhez egy „*iView X Hi-Speed System*” típusú szemmozgás-mérő berendezést használtam. A „training-set”, azaz a videó, amit az alanyok a paraméterek *beállításához* megnéztek, egy 512x298 pixel/frame, 96 dpi-s, 267 frame-ből álló, 8 fps-os (~33 mp), hang nélküli, négy jelenetből álló, természeti képeket (madarak, hegyek, tavak, lovak, stb.) tartalmazó videó volt. A természeti képek, jelenetek használata azért volt célszerű, mert irodalmi adatok szerint ilyen inger esetén – amennyiben az alanynak

nincs egyéb speciális feladata (mint pl. „Hány piros és hány kék papagáj van a képen?”, „Melyik kontinensre tenné a látott tájat?”) a vizuális figyelem nagyrészt bottom-up vezérelt.

A mérések során minden egyes csatornára 40 különböző méretű receptív mező hatékonyságát vizsgáltam meg. Látószögben kifejezve ezek kb. 0.5°-tól ~26°-ig terjedtek.

A csatorna-súlyok meghatározásához különböző hipotéziseket alkalmaztam arra vonatkozóan, hogy adott ingerre (videó adott frame-jére) mely csatornák vehetnek részt a szakkád kiváltásában. (*Szakkád*nak azokat a kis szem-mozgásokat, „ugrásokat” nevezzük, amelyek során a fixáció pozíciója megváltozik, tehát amikor a tér egy újabb pontja kerül az éleslátás középpontjába a régi helyett.[♣])

A mérések során 240 Hz-es mintavételi frekvenciát alkalmaztam, és csak a legalább 1°-os szakkádokat vettem figyelembe.

A mért adatok kiértékelésére a különböző hipotéziseknek megfelelő (mikor mely csatornák tekinthetők a szakkád kiváltójának) kiértékelő programokat MatLab alatt írtam.

Az így meghatározott paraméterek *validációjához* szintén emberi szemmozgás-méréseket végeztem, ugyanazzal a készülékkel és hasonló témájú videón (természeti jelenségek). A többi beállítás megegyezett, de a pontosság kedvéért itt egy hosszabb ingert

[♣] A „*szakkád*” kifejezéssel a szakirodalomban találkozhatunk olyan értelmezésben is, hogy az az egész látótér egységes elmozdulását jelenti, de a jelen dolgozatban én végig a szövegben meghatározott értelemben használom.

használtam: ez 9 jelenetet tartalmazott, a 447 frame ~56 másodpercet vett igénybe.

A validációs eljárás során azt vizsgáltam, hogy a modellem mennyire egyezik az emberek szemmozgás mintázatával. Minden frame-re meghatároztam több pontot, mint valószínű fixáció-pozíciókat. (Mint: "ezen a frame-en a legvalószínűbb fixáció-pozíció *x-y* koordinátája ez és ez, a második legvalószínűbb fixáció-pozíció *x-y* koordinátája ez és ez, stb.) Az eredmények igen jók voltak: az esetek 70%-ában a *mért* pozíció egybeesett az első négy *előre jelzett* lokáció valamelyikével. Ennek a véletlen valószínűsége kisebb, mint 20%.

Vizuális figyelmi modell kialakításakor az a célunk, hogy minél jobban 'lemásoljuk' az élőlények azon képességét, hogy valós időben képesek kiszűrni a számukra éppen fontos információt az érdektelen adattömegből. Ehhez lehet heurisztikus módszereket, ötleteket alkalmazni, de a végső cél – főleg neuromorf modellezés esetén – a mintaként használt élőlény vagy élőlénycsoport idegi mechanizmusainak pontos megismerése és mesterséges eszközön való megvalósítása. Ez nem csak azért fontos, hogy a „biológiailag inspirált" vagy „neuromorf" jelzőket a modellünkre alkalmazhassuk. Igen jelentős szempont az is, hogy ilyen modellek alakítása közben *rengeteget* lehet tanulni az élő rendszerek működéséből, és fölvetődhetnek olyan kérdések is, amelyek aztán a biológiai kutatásokra is visszahathatnak – ami egyébként a tudományterületek közötti párbeszédet is elősegíti. Másrészről pedig, hatékonyság

tekintetében a heurisztikus rendszerek aligha veszik fel a versenyt az élő rendszerek megfelelő mechanizmusaival.

3.Új tudományos eredmények

I. Téziscsoport: Egy új, hatékony módszer a bottom-up vizuális figyelmi modell kialakításában. Az új biológiai eredményeken alapuló, több-csatornás emlős retina modell használata a heurisztikus alacsony szintű vizuális tulajdonság-szűrés helyett; Ennek következményei.

Élő rendszerekben már a retinában elkezdődik az információfeldolgozás. Mi több, a retinából már több csatornán, erősen szűrt és rendszerezett módon távozik az információ, és megy a magasabb agyterületek felé további feldolgozásra. Ezen információ-osztályozás első pontosabb neurobiológiai leírásai – és ennek következtében a rájuk épülő retina-modellek első képviselői is – csak az elmúlt pár évben jelentek meg. Az eddigi figyelmi modellek ezt a retinális feldolgozást egyáltalán nem veszik figyelembe, hanem helyette egy heurisztikus, alacsony szintű (vagy más terminológiával élve: "lokális") vizuális tulajdonság szűrést alkalmaznak.

A modell fő újdonsága egyrészt ezen emlős retina-modell használata – felhasználva ezáltal a retina kutatás legújabb eredményeit –, másrészt az ezen csatornákhoz tartozó receptív mező

méretek meghatározása, és a vonatkozó feltűnőségi térképek eszerinti kialakítása.

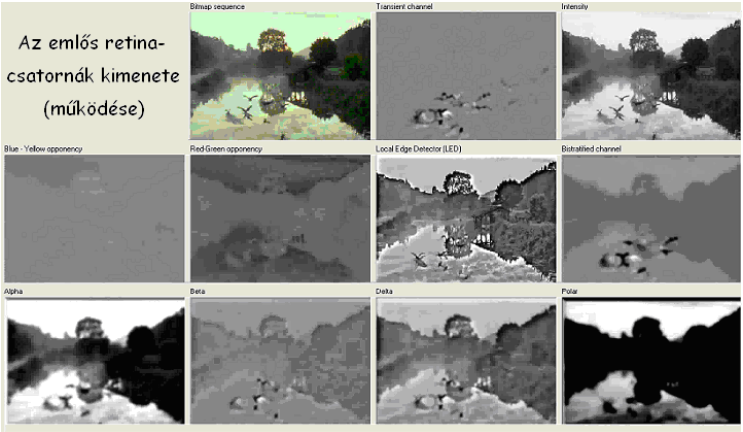
I.1. A klasszikus, (heurisztikus szűréseket alkalmazó) vizuális figyelmi modellt továbbfejlesztettem oly módon, hogy abban az általában használt 3-5, alacsonyszintű (vagy "lokális") szűrés helyett az új biológiai eredményeken nyugvó többcsatornás emlős retina modellt alkalmazom.

A neuromorf vizuális figyelmi modellek első lépése a bemenő kép/videó felbontása, szűrése ún. „alacsony szintű vizuális tulajdonságok" szerint (1. ábra). A jelenlegi modellek ebből jellemzően 3-5 félét alkalmaznak, úgy, mint pl. él-szűrés, sarok-szűrés, színszűrés, stb.

Modellemben e helyett a nemrég felismert és modellezett emlős retina hálózati modellt alkalmaztam, amely tíz csatornát különböztet meg (2. ábra). Ezek közül ötnek a funkciója leolvasható a csatorna kimenetekről (él-szűrés, mozgás-szűrés, intenzitás és két szín opponencia csatorna)[*], a maradék öt funkciója azonban olyan értelemben ismeretlen, hogy működésük célját szavakkal - mindeddig legalábbis - nem sikerült leírni,

[*] Újabb kutatási eredmények szerint, a retina bizonyos sejtjei mozgásra irány-szelektíven reagálnak, vagyis bizonyos élőlényekben létezhet egy irány-szelektíven mozgást szűrő csatorna is. (Fried, S. I., Muench, T. A., & Werblin, F. S. (2002). Mechanisms and circuitry underlying direction selectivity, in the retina. Nature, 420, 411-414.)

megfogalmazni. Ennek egyik következménye az, hogy heurisztikus csatorna-modellekbe elvileg sem kerülhettek be.



2. ábra: Az emlős retina-csatornák működése mozgókép bemeneten. Bal oldali kép: repülő madarak egy tó fölött. Mellette jobbra a „Tranziens" csatorna, ami mindent, ami *mozog* a bemeneten kiszűr, 'megtalál'. Jelen esetben a repülő madarak váltanak ki választ. A környezet azon részére, ahol nincs mozgás, ez a csatorna egyáltalán nem reagál. Tőle jobbra: „Intenzitás", középső sor balról jobbra: kék-sárga szín opponencia csatorna, piros-zöld szín opponencia csatorna, LED (mint „Local Edge Detector", élszűrés) és a „Bistratified" csatorna kimenete látható. Ez utóbbi funkcióját, csakúgy, mint az alsó sorban szereplő csatornákét, eleddig még nem sikerült megfogalmazni. Alsó sor balról jobbra: „Alfa", „Béta", „Delta" és „Polar".

Következmény: Modellemben a fixációs pont meghatározásában az ismeretlen funkciójú csatornákra épülő feltűnőségi térképek is szerepet kapnak – mint ahogy az élőlények idegrendszerében is. Ezek szerepét mozgókép bemenettel vizsgáltam.

Az ún. „feltűnőségi térképek" a fizikai világnak olyan két-dimenziós topografikus leképezései az agyban, amelyekben az

egyes neuronok aktivitása azzal arányos, hogy a külvilág megfelelő pontja mennyire feltűnő, mennyire üt el a környezetétől.

Mivel az előző pontban említett ismeretlen funkciójú csatornákra is épülnek feltűnőségi térképek, amelyek azután részt vesznek a „végső" figyelmi térkép kialakításában, figyelmen kívül hagyásuk jelentősen módosíthatja a végeredményt. Modellemben az összes csatornára épülő feltűnőségi térképet figyelembe vettem, és azok súlyát, „fontosságát" ugyanolyan módszerekkel meghatároztam, mint az ismert funkciójúakét.

A tíz csatornából hétnek (Tranziens, LED, Bistratified, Alfa, Béta, Delta, Polar) a válasza függ az inger időbeli tulajdonságaitól is, azaz kimenetük nem csak az aktuális ingertől, hanem annak előzményeitől is függ. Másképpen fogalmazva: ezek a csatornák - és így a rájuk épülő feltűnőségi térképek is -, többé vagy kevésbe, de reagálnak a *változásokra, mozgásokra*. Méréseim során először került vizsgálatra az ezen csatornákra épülő feltűnőségi térképek hatása a bottom-up vizuális figyelem kialakításában.

II. Téziscsoport: ***A modell ismeretlen paramétereinek – az egyes retina-csatornákhoz tartozó receptív mező méreteknek, illetve a csatornasúlyoknak – a meghatározása, optimalizálása emberi szemmozgás-mérések alapján; A modell verifikálása ugyancsak humán mérések alapján.***

A modell tartalmaz két igen fontos, de ismeretlen paramétert: egyrészt, hogy az egyes retinacsatornák milyen (mekkora) receptív mezővel alakítják ki a hozzájuk tartozó feltűnőségi térképeket, a másik pedig, hogy ezek a térképek aztán milyen súlyozás mellett hozzák létre a végső, „final" feltűnőségi leképezést. (Ezeket az első ábrán piros kérdőjelekkel jelöltem.) A kérdéses értékekre humán szemmozgás-mérések alapján következtettem, majd hasonló mérésekkel ellenőriztem az így beállított modell pontosságát.

Közvetlenül mérni csak a fixációs pontokat tudjuk, ahova a kísérletben résztvevő alanyok néztek, az egyes csatorna-függő feltűnőségi térképeket (amelyek tehát az egyes retina csatornákhoz tartoztak), illetve azok hatékonyságát: *nem*. Ezekre csak *közvetett módon*, különböző feltevésekkel, megközelítésekkel tudunk *következtetni*. Ugyanez igaz a csatornafüggő térképek súlyozott összegére is, az ún. „final" vagy „master" leképezésre. (Igen nehéz olyan kísérletet, „ingert" tervezni, amire kizárólag egyetlen csatorna reagálna – gondoljunk csak arra, hogy ha pl. az inger *mozog*, akkor már rögtön mind a hét időfüggő csatorna ad választ, az Intenzitás csatorna is mindig reagál, stb.)

Mivel irodalmi adatok szerint a bottom-up figyelmi mechanizmus által irányított szemmozgást gyakorlatilag ezek a feltűnőségi térképek (vagy leképezések) határozzák meg, a feltűnőségi térképek *hatékonyságát* azon keresztül határoztam meg, hogy azoknak az erősen feltűnő, azaz nagy intenzitású/értékű pontjai mennyire esnek egybe a mért (tehát emberi) fixáció-pozíciókkal.
Így tehát a beállítandó paramétereket *következtetések* útján kaptam – ezért különösen is fontos a beállított rendszer ellenőrzése, „validációja", emberi szemmozgással való összevetése.

I.1. A modell mind a tíz retina-csatornájára, humán mérések alapján meghatároztam az optimális receptív-mező (RF) méretet, amely szerint az adott csatorna a saját feltűnőségi térképét kialakítja. Ez 40 különböző RF méret megvizsgálását jelentette, látószögben kifejezve ~0.5°-ostól ~26°-osig.

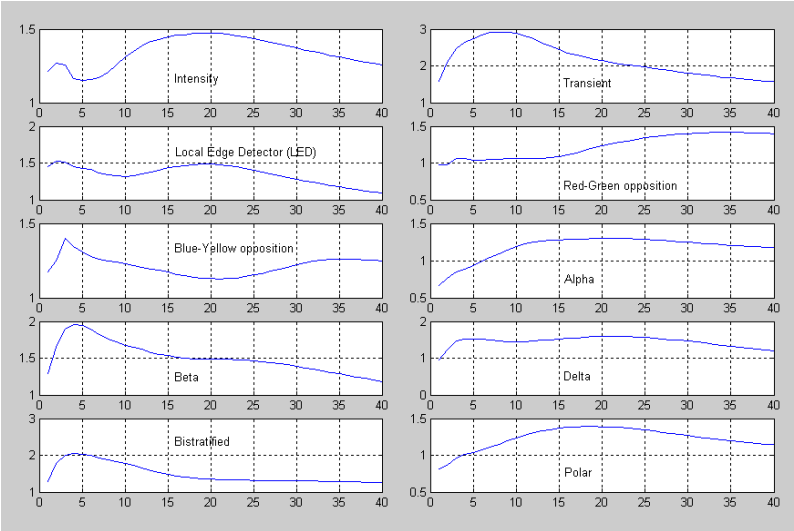
Ugyanarra a bemenetre, különböző receptív mező mértek mellett különböző feltűnőségi térképet kapunk. *Optimálisnak* azt a receptív mező méretet tekintem, amelyhez tartozó térkép a *leghatékonyabb*, azaz amelynek a nagy értékű pontjai a leginkább egybeesnek a *mért* fixáció-pozíciókkal.

A különböző szakkádokat más-más csatornák idézik elő. Mind az, hogy *hány* darab csatorna vált ki egy szakkádot, és hogy ezek

pontosan *melyek*: kérdéses. Ezek meghatározására két különböző feltevést vizsgáltam meg:

1) azok a csatornák váltják ki az egyes szakkádokat (határozzák meg az új fixáció-pozíciókat), amelyek nagyon „hatékonyak" *valamilyen tetszőleges* (egy vagy csak néhány) receptív mező méret mellett.
2) azok a csatornák idézik elő az egyes szakkádokat, amelyek *átlagban* hatékonyak, „feltűnőek", tehát az átlagolásban az *összes receptív mező méret* szerinti feltűnőségi leképezés részt vesz.

Megvizsgáltam az eredményeket, ha az első 1, 3 és 5 'leghatékonyabb' csatornát (mindkét megközelítés szerint) vesszük figyelembe a végső feltűnőségi leképezés kialakításakor.
Az egyes esetek vizsgálata során a 3. ábrához hasonló görbéket kaptam. Ez az ábra a legpontosabbnak bizonyult közelítéshez tartozó görbéket mutatja.



3. ábra: Mind a tíz kis ábra egy-egy csatorna átlagos feltűnőségi („saliency") értéket mutatja a 40 db különböző receptív mező méret szerint meghatározva. Minden egyes képkockára, minden csatornára és receptív mező méretre meghatároztam az átlagos feltűnőségi értéket, és *feltettem*, hogy azok a csatornák vesznek részt az egyes szakkádok kiváltásában, amelyek a legnagyobb *átlagos* feltűnőségi értéket adják az adott ingeren (képkockán). Az első három átlagosan „legfeltűnőbb" csatorna eredményét mutatja az ábra.

Az egyes csatornákhoz tartozó '*optimális'* receptív mező mértek azok, ahol az egyes görbéknek maximumuk van (I. táblázat). A végső, 'behangolt' modellben ezen receptív mező méretekkel lettek a csatorna-függő feltűnőségi térképek meghatározva.

„Valódi", *élő* retinában, az egyes csatornák többféle receptív mezőt is tartalmaznak. Biológiai megközelítésben, a fenti görbék valószínűleg ezek *eloszlását* (sűrűségét) közelíti. Ennek ellenőrzése azonban nem volt a kutatásom tárgya, sőt,

hangsúlyozni szeretném, hogy a fenti „modell-szintű” mérésekből neurobiológiai következtetések nem vonhatóak le.

Int	Tr	LED	R-G	B-Y	α	β	δ	Bist	Pol
20	9	21	20	31	22	19	4	15	15
12,9°	5,5°	13,6°	12,9°	20,2°	14,2°	12,2°	2,18°	9,6°	9,6°

I. táblázat: Az egyes retina-csatornákhoz tartozó optimális receptív mező méretek. Az első sor a csatorna nevek rövidítése, amelyek sorrendben a következők: 1)„Intensity”, 2) „Tranziens”, 3)LED (Local Edge Detector), 4)Piros-zöld szín-opponencia csatorna, 5)Kék-sárga szín-opponencia csatorna, 6)”Alfa”, 7)”Béta”, 8)”Delta”, 9)„Bistratified”, 10)„Polar” A középső sor az optimális receptív mezők *indexei* (3. ábra), míg az alsó sor ugyanez *látószögben* kifejezve. Az index és a látószög közötti összefüggés: $\mathrm{tg}\frac{\alpha}{2} = \frac{4i-3}{100} * 0{,}147$, ahol *i* az indexet jelenti.

I.2. Különböző hipotéziseket vizsgáltam meg arra vonatkozóan, hogy az egyes csatornák milyen arányban, *milyen súllyal* felelősek a szakkádok kiváltásában, az új fixáció-pozíciók meghatározásában. Ezek alapján különböző csatorna-súlyozásokhoz jutottam.

Megvizsgáltam több olyan hipotézist, amelyben az egyes csatorna-súlyok állandóak, vagyis a csatorna-függő feltűnőségi leképezések mindig ugyanolyan arányban vesznek részt a „final” feltűnőségi térkép kialakításában, és megvizsgáltam olyan hipotéziseket is, amelyben a csatornák súlya az ingertől függően folyamatosan változik.

- A fix csatorna-súlyozási feltevések - amelyek erősen épülnek az előző pontra -, a következők voltak:

Az egyes csatorna-súlyok azzal arányosak, hogy azok az előzetes mérések során *milyen arányban*, (milyen százalékban) bizonyultak szakkád-kiváltónak…

i. …*tetszőleges* receptív mező méret szerint

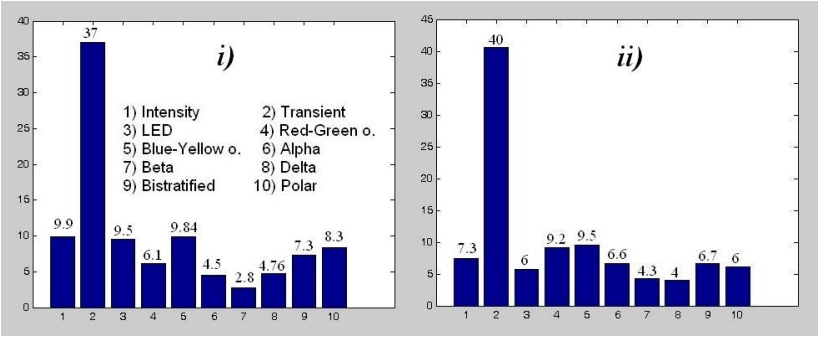
(Vagyis az egyes csatornák súlya azzal arányos, hogy a hozzájuk tartozó feltűnőségi térképek milyen gyakorisággal tartalmazták a legnagyobb értéket a mért fixációs pontokban.)

ii. …*átlagos* (tehát az összes receptív mező méret szerint számolt) feltűnőségi érték szerint

(Más szóval, a mért fixációs pontokban minden csatornára kiszámoltam az összes receptív mező szerinti „feltűnőségi értékeket", azokat átlagoltam, és a csatornasúlyok ezen átlagok szerint alakultak.)

Az eredményeket a 4. ábra mutatja, míg az egyes megközelítések pontosságát az 5 ábra.

(Az *i)* és az *ii)* jelölések a fenti két megközelítésre vonatkoznak.)



4. ábra: A becsült csatorna-súlyok az egyes fix súlyozási megközelítések szerint. A becslés elve: melyik csatorna hányszor bizonyult szakkád-kiváltónak a különböző megközelítések szerint. Az egyes hipotézisek pontosságát az 5. ábra mutatja.
(Érdemes megfigyelni, hogy az *y* tengelyen a beosztás nem ugyanaz.)

- A változó („adaptív") csatorna-súlyozási hipotézisek azon a feltevésen alapulnak, hogy az, hogy éppen *melyik* csatorna, és az adott csatorna *milyen súllyal* vált ki egy szakkádot, az az ingertől függ. Más szóval, az aktuális csatorna-súlyok inger-függőek, és nem előre definiáltak.

Ebben a megközelítésben az egyes csatorna-súlyok azzal arányosak, hogy az *adott (aktuális) ingeren, képkockán*, a mért fixáció pozícióban mennyire „feltűnőek" …

i. …*tetszőleges* receptív mező méret szerint
ii. …a negyven különböző RF méret által meghatározott feltűnési értékek *átlaga* szerint

(hasonlóan az előzőekhez).

A várakozásokkal ellentétben – habár a különbség kicsi volt -, ezek a fix súlyozású stratégiák bizonyultak jobbaknak olyan

értelemben, hogy ezek pontosabb előrejelzéseket adtak az emberi fixáció pozíciókra. A fix súlyozású hipotézisek kb. 5%-kal bizonyultak pontosabbnak ezen megközelítéseknél.

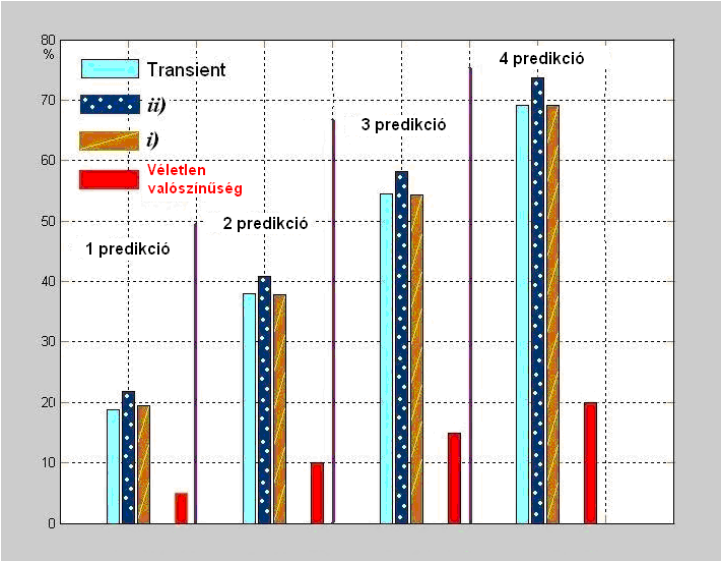
Validáció: A modell pontosságát emberi szemmozgás-mérésekkel igazoltam, és megmutattam, hogy komplex természeti képeken a modell igen jól előrejelzi az emberi fixáció-pozíciókat.

A fenti mérések alapján beállított modellel predikciókat, várható fixáció-pozíciókat adtam meg, amiket azután összehasonlítottam humán szemmozgás mérésekből származó adatokkal. A *mért* fixáció-pozíciók ~70%-ban esetek egybe az adott képkockán az első négy legvalószínűbbnek *predikált* hely valamelyikével (- a pontos értékek az egyes megközelítések szerint kicsit változtak). Ennek a véletlen valószínűsége kisebb, mint 20%. „Találatnak" minősítettem azt, ha az *előre jelzett* és a *mért* pozíció közötti távolság 5°-nál kisebb volt.[♦]

Az 5. ábra a fix csatorna-súlyozású megközelítések pontosságát mutatja. A két szövegben szereplő megközelítést kiegészíti egy harmadik is, amikor végig egyetlen csatornát vettem csak figyelembe: a mozgást szűrő „Tranziens" csatorna feltűnőségi térképét egyben „final" leképezésként definiáltam. Ez azért volt

[♦] 10°-kal számolva a találati pontosság nagyon megnő – habár nyilván a „véletlen valószínűség" értéke is. Az 5°-os 'határ' mind biológiai, mind kiértékelési szempontból ésszerű választásnak tűnt.

érdekes, mert irodalmi adatok szerint dinamikus bemeneten ez a csatorna rendkívül erős – ami intuitíven persze nem meglepő, pl. ha arra gondolunk, hogy akár látóterünk perifériáján megmozduló macska, madár, stb. hatására milyen reflex-szerűen odakapjuk a tekintetünket. Ezt az eredményt az én méréseim is alátámasztják.



5. ábra: A fix csatorna-súlyozású megközelítések pontossága. Az első oszlop-hármas azt mutatja, hogy az egyes megközelítések szerint, a *mért* figyelmi fókuszok hány százalékban estek egybe az 1 db, legvalószínűbbnek *előre jelzett* fixáció-pozíciókkal. Mellettük a piros, csíkkal jelzett oszlop ennek a véletlen valószínűségét mutatja.
Az utolsó oszlop-trión azt láthatjuk, hogy ha 4 valószínű figyelmi fókuszt adunk meg, akkor a mért pozíció hány százalékban esett ezek valamelyikével egybe. Leolvasható, hogy ez 70% körül mozog az egyes esetek szerint. Ugyanennek a véletlen valószínűsége kisebb, mint 20%.
Hasonlóan, a második és a harmadik oszlop-hármas az ahhoz tartozó eredményeket mutatja, hogy ha kettő-, illetve három valószínű fixáció-pozíciót adunk meg.
Érdemes megjegyezni, hogy két különböző ember jó eséllyel más-más helyre fixál ugyanazon a képkockán, illetve ugyanaz az alany is valószínűleg máshova fog nézni, ha többször nézetjük meg vele ugyanazt a videót. Így az erősen javuló tendencia nem meglepő.

4.Az eredmények felhasználási területe

Figyelmi modellek alkalmazási területei rendkívül változatosak, a bennük felhasznált módszerek és részfeladatok rengeteg területen felhasználhatóak. Így az elmúlt években a fent leírt modell különböző részeit nekem is volt alkalmam kipróbálni konkrét gyakorlati alkalmazásokban is - egész pontosan az ún. Bionikus Szemüveg Project egyes részfeladataiban.

Ez a project a látáskárosultak mindennapi életét hivatott segíteni mobil eszközzel, folyam elemzés és felismerés segítségével. A főbb irányvonalakat, csakúgy, mint az alfeladatokat, a Vakok és Gyengénlátók Országos Szövetségének szakértőjével együtt határoztuk meg. Ezen project keretében a fenti modellt, illetve annak egy elő-feldolgozó résszel kibővített változatát, amely az instabil bemeneti videót egy speciálisan ide tervezett algoritmussal stabilizálja, már sikeresen illesztettem több, a projectben meghatározott alfeladathoz. Ezek a videó-bemenetek általában rendkívül zajosak, a nagy és kiszámíthatatlan kameramozgások miatt az egyes frame-ek erősen bemozdultak, továbbá gyakori az egyik képkockáról a másikra való nagymértékű elmozdulás is, pl. amikor a felhasználó befordul egy sarkon. A stabilizáló elő-feldolgozó rész célja, hogy az álló objektumokat (pl. épületek) ugyanazon a pixelpozíción tartsa, míg a mozgó részek (pl. járókelők) pixelpozíciója változhat.

Ennek során az alapötlet egy „optic-flow” algoritmus és egy affin transzformációs modell ötvözése volt, amely kezelni tudja a haladás, elfordulás, kicsinyítés-nagyítás („skálázás”), illetve a különböző

vetületek okozta torzulásokat. Az optic-flow algoritmus használatával az egyes *pixelek sebessége becsülhető*, azok időbeli- és térbeli (vertikális és horizontális) gradienseinek mérésével. Majd ebből a *frame* elmozdulása (függőleges és vízszintes irányban), elfordulása, skálázása és vetülete határozható meg a transzformációs modell segítségével.

A több-csatornás emlős retina modell segítségével kikerülhető (legalábbis részben) az egyik legnagyobb probléma, amivel napjaink képfeldolgozó algoritmusai szembesülnek, nevezetesen az, hogy egy objektum tetszőleges pontjának színértékei vagy intenzitás értéke nagyon erősen függ az aktuális megvilágítási körülményektől. Ez alapvető fontossággal bír a különböző gyakorlati alkalmazások tervezésénél, mint ahogy az alábbi, a Bionikus Szemüveg Projectben felmerülő részfeladatok megoldása során is központi elem volt az említett retina csatorna felbontás. Ezen részfeladatok a következők voltak:

- LED-kijelzők megtalálása (beltéri vagy kültéri, valós életből vett helyszíneken)
- Közlekedési jelzőtáblák lokalizációja

  Ennél a két feladatnál a fő cél egy olyan algoritmus készítése, amely a bemeneti képen gyorsan képes lokalizálni azokat a részeket („ROI”-kat / „Region of Interest”), ahol nagy valószínűséggel LED indikátorok illetve közlekedési táblák vannak. Így egy klasszifikáló algoritmusnak csak ezen részeket kell feldolgoznia a teljes bemenet helyett, ami

lényegesen felgyorsíthatja a teljes folyamatot. A főbb nehézségek az amúgy is rossz felbontású bemenet bemozdulásából és *sokféleségéből* adódtak (olyan értelemben, hogy nem lehetett felkészülni pl. „tipikus" helyszínekre vagy körülményekre).

Az algoritmusok pontossága 80% körül van. A teszt adatbázisok komplex, valós életből vett helyszíneket tartalmaznak.

- Elsődleges fényforrások (pl. lámpák) detektálása.

  Ez a feladat – habár egy átlagos látással bíró személynek triviálisnak tűnik – sok kellemetlenségtől kímélhet meg egy gyengénlátó vagy vak embert, például azáltal, hogy a lámpák nem maradnak teljesen fölöslegesen égve hetekig egy vendég után. Itt az a legfontosabb kritérium, hogy az algoritmus biztosítson megvilágítás-független megoldást, vagyis az eredmény legyen egyformán megbízható egy sötét cellában és egy fényben úszó nyári szobában.

  Az algoritmus egyetlen retina-csatornát használ fel (nevezetesen a „Polar" csatornát), amelyre rendkívül megbízható alkalmazás építhető: a helyes válaszok aránya 99%.

A fenti alkalmazásokról és megoldásokról külön publikációkban számoltam be.

Általánosságban, egy jól működő vizuális figyelmi rendszer lehetséges felhasználási területe szinte felsorolhatatlanul széles, kezdve a térfigyelő rendszerektől a robot-látáson át a különböző 'bionikus' alkalmazásokig. Egy jól működő *bottom-up* rendszer (amelynek létrehozására doktoranduszi éveim alatt törekedtem) azonban távolról sem *teljes* figyelmi rendszer. Teljes akkor lenne, ha az ún. „top-down" mechanizmust is tartalmazná. Erről a kérgi eredetű modulációról azonban jelenleg még igen keveset tudunk, annál mindenesetre biztos kevesebbet, hogy az egészet, úgy ahogy van, megbízhatóan modellezni tudjuk.

Azonban némi ismeretünk van már: ismert irodalmi adat például, hogy ez a mechanizmus a csatorna-súlyozás módosításánál van „visszakapcsolva" a bottom-up körbe, közvetlenül a „final" feltűnőségi térkép kialakítása előtt. (1. ábra, alul középen) Ennek mintájára sok gyakorlati alkalmazás készíthető például ezen súlyozások alkalmazás-függő módosításával (pl. a fent említett algoritmusok közül a LED kijelzők vagy közlekedési táblák megtalálása).

5.Köszönetnyilvánítás

Elsősorban köszönetet szeretnék mondani Roska Tamás professzor úrnak rengeteg támogatásáért, útmutatásáért és segítségéért, a lelkesítő beszélgetésekért és konzultációkért, és mert mindig időt tudott rám szakítani megannyi teendője mellett is.

Köszönöm konzulensemnek, Vidnyánszky Zoltánnak is a velem való törődést, a jó-tanácsokat és beszélgetéseket, a cikkeim gondos

átolvasását, és nem utolsó sorban, hogy rajta keresztül egy kis bepillantást nyerhettem a biológusok világába, gondolkodásmódjába.

Köszönöm évfolyamtársaimnak, Bárdi Tamásnak, Hillier Daninak Havasi Lacinak, Kóbor Istvánnak és Vásárhelyi Gábornak a sok segítséget, akármiről volt is szó, kezdve a számítógépektől a baráti tanácsokig, és köszönöm ugyanezt Cserey Gyurinak, Iván Kristófnak, Soós Gergőnek, Hegyi Barnának, Harcos Tamásnak Benedek Csabának, Bankó Évának és Gál Viktornak is.

Külön köszönet Bálya Dávidnak az első években nyújtott aktív segítségért és a retinamodellért, amelyet doktoranduszi éveim alatt végig használtam.
Hasonlóképp meg szeretném köszönni a „látásos-" avagy „bionikus szemüveg"-csoport tagjainak az együttműködést, segítséget és beszélgetéseket: Bálya Dávidnak, Wagner Robinak és Karacs Kristófnak.

Köszönöm Karmos György professzor úrnak a leuveni utam előtti segítséget, eligazítást és hasznos tanácsokat. Nehezebb lett volna nélkülük!
Köszönöm Marc Van Hulle professzor úrnak, hogy eltölthettem a laborjában, a Leuveni Egyetemen egy rendkívül érdekes és szép emlékű félévet. Csoportjából köszönöm Karl Pauwelsnek a kitartó segítséget, és hogy bármikor, akár órákra is rámért.

Örömmel és hálás szívvel emlékezek vissza Szegeden töltött egyetemi éveimre. Köszönöm Kocsor András ottani témavezetőmnek, hogy első cikkemet, első próbálkozásaimat irányította és segítette.

És végül, de persze távolról sem utolsó sorban köszönök mindent szüleimnek, önfeláldozó szeretetüket és támogatásukat, hogy tanulmányaimban mindig támogattak, hogy mindig megteremtették számomra a tanuláshoz, kutatáshoz szükséges feltételeket is, akár erejükön túl is.

6.Publikációk

Folyóiratokban

[1] **A. Lázár**, Z. Vidnyánszky, T. Roska, "Modeling stimulus-driven attentional selection in dynamic natural scenes," *International Journal of Circuit Theory and Applications*, (nyomtatás alatt)

[2] **A. Lázár,** K. Pauwels, M. Van Hulle, T. Roska, "Scene analysis of unstable video flows – using multiple retina channels and attentional methods," *Integrated Circuits: Research, Technology and Applications*, (elfogadva)

Konferenciákon

[3] **A. K. Lázár**, R. Wagner, D. Bálya, T. Roska, " Functional representations of retina channels via the refineC retina simulator," *Cellular Neural Networks and their Applications. Proceedings of the 8th IEEE international workshop,* pp. 333-338, *2004*, Budapest

[4] Bálya D., **Lázár A.**, " Retinal processing", *XI. MITT Kongresszus*, Pécs (*2005*)

[5] Vidnyánszky Z., Kovács G., **Lázár A.**, "Active vision" , *XI. MITT Kongresszus*, Pécs (*2005*)

[6] **A. Lázár**, A. Kocsor, " An application of ranking methods: retrieving the importance order of decision factors," *IEEE International Workshop on Soft Computing Applications SOFA 2005*, Szeged, Hungary – Arad, Romania

[7] T. Roska, D. Bálya, **A. Lázár**, K. Karacs, R. Wágner, M. Szuhaj, "System aspects of a bionic eyeglass", *Proc. of International Symposium on Circuits and Systems ISCAS,* pp. 161-164, *2006*, Kos, Greece

[8] **A. Lázár**, T. Roska, "Human Tested Saliency Map Generation in the Bionic Eyeglass Project", *Proceedings of The 10th IEEE International Workshop on Cellular Neural Networks and their Applications,* pp.91-95, *2006*, Istanbul, Turkey

[9] K. Karacs, **A. Lázár**, R. Wagner, D. Bálya, T. Roska, "Bionic Eyeglass: an Audio Guide for Visually Impaired," *Proceedings of the 1st Biomedical Circuits and Systems Conference,* pp. 190-193, *2006*, London, UK

7.A témához kapcsolódó irodalom

[1] L. O. Chua, T. Roska, "Cellular Neural Networks and Visual Computing", *Cambridge University Press*, Cambridge, UK, 2002.

[2] F. S. Werblin, T. Roska and L. O. Chua, "The analogic cellular neural network as a bionic eye," *Intl. J. of Circuit Theory and Applications*; Vol. 23, pp. 541-569, 1995

[3] B. Roska and F. S. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature*, Vol. 410, pp. 583-587, 2001.

[4] D. Bálya, B. Roska, T. Roska, F. S. Werblin, "A CNN Framework for Modeling Parallel Processing in a Mammalian Retina," *Int'l Journal on Circuit Theory and Applications*, Vol. 30, pp. 363-393, 2002

[5] L. Itti, "Modeling Primate Visual Attention," **In:** *Computational Neuroscience: A Comprehensive Approach,* (J. Feng **Ed.**), pp. 635-655, Boca Raton: CRC Press, 2003

[6] L. Itti and Christof Koch, "Computational modeling of visual attention," *Nature Neuroscience,* Vol 2, 2001

[7] Richard H. Mashland "The fundamental plan of the retina", *Nature neuroscience* Vol 4 No. 9, 2001

[8] E. R. Kandel and J. H. Schwartz, "Principles of Neuroscience" Elsevier, New York, 3rd edition, 1991

[9] D. J. Parkhurst, E. Niebur „Stimulus-driven guidance of visual attention in natural scenes” **In** Neurobiology of attention, (L. Itti, G. Rees, J. K. Tsotsos **Ed.**), pp. 240-245, Elsevier, 2005

[10] R. Carmi, L. Itti „Visual causes versus correlates of attentional selection in dynamic scenes” **In** Vision Research, doi:10.1016/j.visres.2006.08.019 , 2006

[11] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry”, *Hum. Neurobiol*. 4, 219-227, 1985

[12] H. Nothdurft, “Salience from feature contrast: additivity across dimensions”, *Vision Res*. 40, 1183-1201, 2000

[13] S. Shipp, “The brain circuitry of attention”, *Trends in Cognitive Sciences*, Vol.8 No.5, 2004

[14] L. Itti, “Models of Bottom-up Attention and Saliency”, **In:** Neurobiology of Attention, (L. Itti, G. Rees, J. K. Tsotsos **Ed.**), pp. 576-582, San Diego, CA:Elsevier, Jan 2005.

[15] R. Carmi and L. Itti, “The role of memory in guiding attention during natural vision”, *Journal of Vision*, Vol 6, No.9., pp. 898-914, 2006

[16] B. Zitova, J. Flusser, “Image registration: a survey”, *Image and Vision Comp*. Vol 21, 977-1000, 2003