Raport Semestralny

Bartosz Szurgot

20 stycznia 2008

Spis treści

Rozdział 1

Wprowadzenie

Rozdział wprowadzający poświęcony jest pobierznemu przedstawieniu (wprowadzeniu) tematyki omawianej w dalszych częściach opracowania. Pierwsza część jest poświęcona podstawowym pytaniom przedstawianej problematyki – dlaczego warto się nią zajmować oraz skąd wzięła się inspiracja i zapotrzebowanie na tego typu rozwiązania (sekc. 1.1). W drugiej części zostaną zaprezentowane cele jakie naukowcy pragną uzyskać pracując nad omawianą problematyką (sekc. 1.2). Zostaną także zaprezentowane dwa podstawowe nurty jakie można wyróżnić w pracach z dziedziny kierunkowania uwagi.

1.1 Poruszna problematyka

Przykładowe zdjęcie wykonane współczesnym aparatem cyfrowym będzie miało rozdzielczość rzędu $3000x2000$ punktów w 24-bitowej palecie kolorów. Daje to około 6 milionów punktów, z których każdy może mieć jeden z 2^{24} (około 16 milionów) różnych kolorów. Nawet prosta iteracja po tak dużej ilości pikseli jest czasochłonna. Dodatkowo przetwarzanie obrazu składa się zwykle z kilku etapów, przy czym w trakcie każdego z nich obraz jest „przeglądany" kilkukrotnie. Efektem tego jest niemożliwość wykonania praktycznego, użytecznego systemu wizyjnego, pracującego w czasie rzeczywistym (ang. *RT – real-time*), na tak dużych obrazach.

Aby sprostać wymaganiom czasowym i jakościowym potrzebny jest więc etap „preselekcji" wybranych fragmentów obrazu wejściowego [46], tak aby ilość informacji jaką trzeba przetworzyć była jak najmniejsza. Jednocześnie, w procesie eliminacji, nie można tracić elementów istotnych z punktu widzenia docelowej analizy.

Z efektami skupiania uwagi spotykamy się na codzień, kiedy przykładowo widzimy coś, jednocześnie niezauważając tego. Dzieje się tak gdy obiekt jest w polu naszego widzenia, jednak jest poza naszą percepcją – nasza uwaga jest aktualnie skoncentrowana na innym obserwowanym elemencie. Zjawisko to pokazano w 1970 r. w eksperymencie nazwanym niezauważaniem zmian (własne tłumaczenie z ang. *change blindenss*) [16]. Przykładem sytuacji kiedy owe zjawisko zachodzi jest naprzemienne „mruganie" dwóch obrazów rozdzielonych krótką (np: $200[ms]$) pustą planszą. Wynikiem tego jest niezauważanie przez obserwatora istotnych zmian na obrazie. Można obejrzeć przykładową aplikację pokazującą to zjawisko na stronie [1].

Mechanizm skupiania uwagi jest bardzo złożony. Ludzie są „wyposażeni" w różne rodzaje mechanizmów koncentracji. Możemy skupić się na obserwowaniu jakiegoś wycinka naszego otoczenia (świadomy wybór), z drugiej jednak stron naszą uwagę zawsze przyciągnie nagły błysk światła lub gwałtowny ruch w polu widzenia [21] (naturalny odruch – reakcja na potencjalne zagrożenie).

W kontekście systemów wizyjnych omawiana problematyka nazywa się ogólnie koncentracją, skupianiem lub kierunkowaniem uwagi[1] [31]. Najczęściej spotykane, anglojęzyczne określenia tejże dziedziny to: *Focuse Of Attention* (bardzo często stosowana jest też skrótowa postać: FOA), *Visual Attention*, *Selective Visual Attention*.

W niniejszej pracy umówione zostaną pojęcia i metody spotykane w tego typu systemach analizujących obrazy rastrowe oraz ich sekwencje. Zdaniem autora, tego typu algorytm powinien być podstawą dla niemal każdego systemu wizyjnego.

1.2 Cele koncentracji uwagi

Dla przykładowego zdjęcia, omawianego w poprzednim podrozdziale ($3000x2000$ punktów, 24-bitowa głębia kolorów) wybranie losowego fragmentu o jakimś z góry ustalonym rozmiarze (np: $100x100$ punktów) co prawda zmniejszy drastycznie czas dalszego przetwarzania (będzie to miej niż 2% rozmiaru początkowego), jednak informacja jaką zyskamy na podstawie takiej analizy prawie na pewno będzie szczątkowa i prawdopodobnie bezużyteczna. Jeżeliby jednak wybrać fragment (lub fragmenty) w sposób „przemyślany", maksymalizując ilość informacji jaką niesie, jednocześnie minimalizując jego rozmiar, możnaby uzyskać znaczące (nawet kilkunastorkotne) skrócenie czasu

[1] Wymienione nazwy będą stosowane przez autora zamiennie.

całkowitej analizy, przy minimalnym zmniejszeniu dokładności uzyskanych w ten sposób wyników.

Choć rozwiązanie takie wydaje się być jedynie pewnym chwytem mającym na celu uproszczenie obliczeń wykonywanych przez komputer, problem sięga głębiej. Badania nad ludzką percepcją wykazały jednoznacznie, iż człowiek nie jest w stanie „analizować" wszystkiego co widzi w danej chwili. Przeciętny osobnik potrafi śledzić około 4 niezależnych punktów przez kilkanaście sekund [39] zaś szukanie na obrazie określonego obiektu, nie posiadającego pojedynczej, wyróżniającej cechy ma złożoność liniową, w zależności od całkowitej ilości obiektów znajdujących się w polu widzenia [27].

Prowadzone są również badania porównujące koncentrację uwagi ludzi oraz istniejących algorytmów [19].

Badania nad koncentracją uwagi można podzielić na dwa podstawowe nurty: „biologiczny" oraz związany z systemami wizyjnymi. W kolejnych podrozdziałach zostaną obrazowo przedstawione oba nurty. Nieco większy nacisk zostanie położony na drugi z prezentowanych nurtów jako iż będzie on przewodni dla tejże pracy.

1.2.1 Nurt „biologiczny"

Celem pierwszego nurtu („biologicznego") jest poznanie i zrozumienie jak działa koncentracja uwagi u ludzi: jakie są jej możliwości oraz ograniczenia. Przedstawiciele tego nurtu zajmują się głównie przeprowadzaniem badań celem poznania zasad działania FOA u zwierząt i ludzi. Czasami tworzone są także modele obliczeniowe mające na celu zweryfikowanie poprawności koncepcji. Zrozumienie jak postrzegamy świat na wczesnych etapach naszej wizji byłoby znaczącym krokiem do zrozumienia jej jako całości.

Najbardziej popularne badania jakie są prowadzone w tej dziedzinie mają na celu rozpoznanie jak funkcjonuje część koncentracji uwagi odpowiadająca za wykrywanie wyróżniających się fragmentów bez stosowania wiedzy dziedzinowej – kierunkowanie uwagi sterowane danymi (ang. *bottom-up*) [27][19]. Eksperymenty takie są jednocześnie obrazowe i proste w wykonaniu.

Nieco mniej popularne są eksperymenty badające kierunkowanie uwagi sterowane wiedzą (ang. *top-down*). Jest jednak wyraźna zależność pomiędzy ukierunkowaniem uwagi a czasem znalezienia poszukiwanego elementu [27] [45] (np: znacznie trudniej jest znaleźć „coś nietypowego na zdjęciu parkingu" niż znaleźć „samochód bez tablicy rejestracyjnej"). Świadczy to jednoznacznie o istnieniu i ogromnym znaczeniu tej właściwości mózgu dla organizmów żywych.

1.2.2 Nurt systemów wizyjnych

Drugi z nurtów stawia jako cel stworzenie podsystemu wizyjnego zdolnego do wybierania istotnych elementów obrazów do dalszego przetwarzania. Choć zdecydowana większość naukowców pracujących nad FOA stosuje podejścia wzorowane na naturze [10][27], zrozumienie działania kierunkowania uwagi u ludzi nie jest celem pierwszoplanowym a jedynie środkiem [45].

Pierwszą istotną koncepcją jaka pojawiła się w FOA dla systemów wizyjnych było stowrzenie koncepcji map występności[2] (ang. *saliency map*) [10]. Zaprezentowany w 1985 r. model ma u podnóża badania biologicznych (naturalnych) systemów wizyjnych, lecz powstał z myślą o implementacji komputerowej. Jest on obecnie powszechnie stosowany w niemal każdej pracy poświęconej kierunkowaniu uwagi. Niektórzy autorzy dodają również pewne rozszerzenia do zestawu używanych cech [36]. Pojawiają się także modyfikacje samego modelu [48][32].

Niezwykle popularne są też podejścia wykorzystujące implementacje oparte częściowo [27] lub w całości [22] o sieci neuronowe. Praktycznie wszystkie prace bazujące na idei map występności posiadają jako część implementacji sieć WTA (ang. *Winner-Take-All*) [18]. Istnieją także prace ukierunkowane na sieci o specyficznych własnościach umożliwiających kierunkowanie uwagi [22][31].

Nie są to jednak jedyne możliwe podejścia [25][43][47]. Dokładny opis działania wybranych spośród nich jest zamieszczony w późniejszych rozdziałach (rozdz. 3).

[2] Inne tłumaczenie to „mapa istotności" [31]. W niniejszym opracowaniu obra określenia będą używane zamiennie.

Rozdział 2

Inspiracja naturą

Rozdział ten prezentuje podstawowe zagadnienia związane z kierunkowaniem uwagi w systemach wizyjnych. Znaczna jego część jest poświęcona badaniom jakie są prowadzone celem obserwacji działania kierunkowania uwagi u ludzi. Uwzględniono także skrótowy opis podziału zadań w przetwarzaniu wizyjnym na poszczególne obszary mózgu (ścieżki „where" oraz „what") oraz zagadnienia symulacji sieci neuronowych na komputerach.

2.1 Badania

W niniejszej sekcji zostanie pokrótce omówionych kilka ważniejszych i ciekawszych eksperymentów jakie można znaleźć w literaturze poświećonej FOA. Zdecydowana większość pochodzi z czasopism skupiających się na takich problemach jak kognitywistyka, psychologia czy sieci neuronowe.

Liczne eksperymenty przeprowadzane przez psychologów oraz badaczy zajmujących się koncentracją uwagi mają swoje odbicie w zmianach modeli/implementacji inspirowanych naturą.

2.1.1 Cechy podstawowe

Na podstawie przeprowadzonych eksperymentów z udziałem ludzi został wyszczególniony zestaw cech jakie są uważane za podstawowe, wykrywane przez ludzki mózg na wczesnych etapach przetwarzania. Na ich podstawie mózg stwierdza czy coś jest warte „przyjżenia się" czy też nie oraz decyduje czy przekazać daną informację dalej do świadomości. Omawiane cechy to [27]:

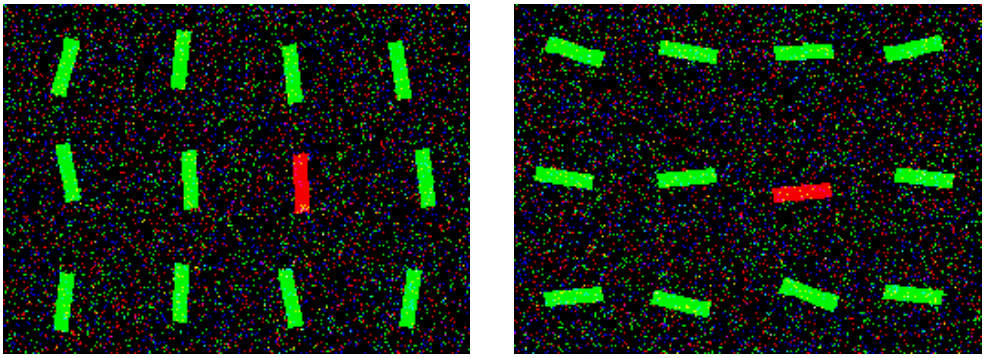
- kolor,
- nachylenie (kąt),

- intensywność (jasność).

Wymienione własności (cechy) mają w mózgu odpowiednik w postaci map informujących jak bardzo dany obszar na widzianym obrazie wyróżnia się względem swojego otoczenia (jak bardzo przyciąga uwagę). Ma to szczególne znacznie dla zjawiska opisanego w podsekc. 2.1.2.

2.1.2 Efekt „wyłaniania”

Istnieje pewien szczególny przypadek, kiedy kierunkowanie uwagi powoduje bardzo wyraźny wzrost wydajności wyszukiwania na obrazach. Jest to sytuacja, w której poszukiwany cel niejako samoistnie „wyłania się” (ang. *pop-out*) z reszty obrazu kierując na siebie uwagę [27][45]. Przykład takiego efektu przedstawiono na rys. 2.1.



(a) (b)

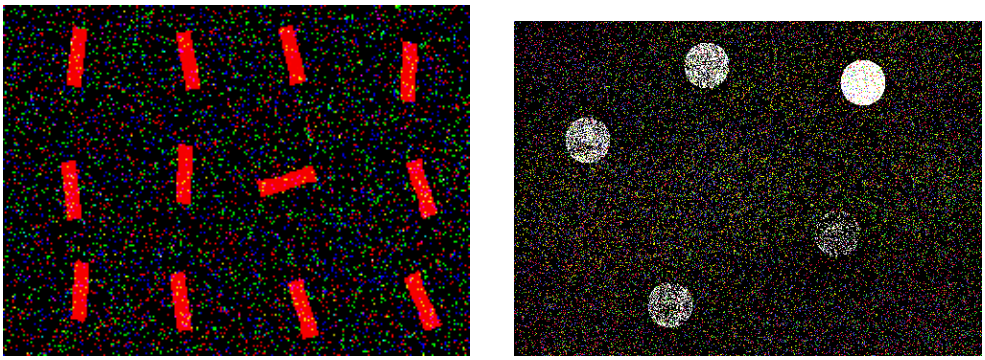
Rysunek 2.1: Efekt „wyłaniania się” zaprezentowany na przykładzie koloru (zaczerpnięte z [2]).

Efekt ten pojawia się kiedy „cel” różni się od wszystkich pozostałych elementów otoczenia minimum jedną cechą podstawową[1] (na rys. 2.1 jest to kolor). Zjawisko to posiada dwie istotne zalety:

1. wyróżnia się istotnie od otoczenia przyciągając uwagę,
2. obiekt wyłaniający się z obrazu jest znajdowany zawsze jako pierwszy („od razu” na początku) niezależnie od liczby elementów otoczenia – tzn: 1 czerwona kreska zostanie znaleziona na obrazie zawierającym 3 zielonych tak samo szybko jak na obrazie zawierającym 30 zielonych kresek [45][27].

[1] Cechy podstawowe są opisane w podsekc. 2.1.1.

Oczywiście podobny efekt będzie występował dla każdej z cech podstawowych. Przedstawiono go dla orientacji (kąta nachylenia) na rys. 2.2(a) oraz dla intensywności na rys. 2.2(b).



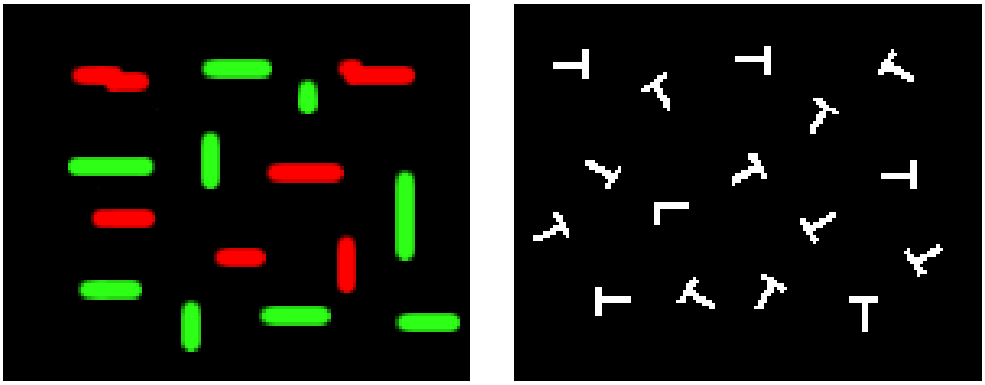
(a) (b)

Rysunek 2.2: Efekt „wyłaniania się” zaprezentowany na przykładzie (a) orientacji (b) intensywności (obrazy zaczerpnięte z [2]).

Jest to bardzo istotny efekt z punktu widzenia kierunkowania uwagi, ponieważ pozwala na natychmiastowe zauważenie pewnych elementów, które człowiek uznałby za „przykuwające uwagę”. Zjawisko to jest powszechnie wykorzystywane na codzień (np: znaki drogowe, ogłoszenia reklamowe, etc… [27]). Kwestia ta będzie jeszcze poruszana przy okazji omawiania stosowanych podejść w rozdz. 3.

2.1.3 Szukanie złożone

Badania pokazały, iż efekt „wyłaniania się” opisany w podsekc. 2.1.2 występuje jedynie w bardzo ściśle określonych przypadkach. W pozostałych sytuacjach kierunkowanie uwagi pomaga wyeliminować niosące mało informacji fragmenty, jednak nie ma wyraźnie jednego „górującego” nad pozostałymi miejsca do odwiedzenia. Takim przypadkiem jest sytuacja, w której trzeba znaleźć pionową, czerwoną kreskę pośród pionowych i poziomych, czerwonych i zielonych kresek. W tym przypadku uwaga jest kierowana kolejno do „interesujących” fragmentów obrazu, lecz znalezienie poszukiwanego celu wymaga liniowego przejścia przez (potencjalnie wszystkie) elementy jakie są na obrazie [27][45]. Przykłady takich obrazów przedstawiono na rys. 2.3.



(a) (b)

Rysunek 2.3: Brak występowania efektu „wyłaniania się” – konieczne jest przeszukiwanie liniowe po elementach „przykuwających” uwagę: (a) kolorowych kreskach i (b) literach (litery zaczerpnięto z [2]).

2.1.4 Niezauważanie zmian

Zjawisko niezauważania istotnych zmian na obrazie, kiedy jest on pokazywany naprzemiennie z obrazem na którym dany element występuje oraz przerywany na krótko pustym obrazem było wspominane w sekc. 1.1. Jest to bardzo ciekawy efekt, który można zaobserwować korzystając z dostępnej w internecie aplikacji na stronie profesora Ronalda A. Rensinka [1].

Choć efekt jest bardzo ciekawy i szokujący nie ma on większego znaczenia z punktu widzenia nurtu koncentrującego się na sztucznych systemach wizyjnych. Tym samym nie będzie on dalej poruszany w niniejszym opracowaniu.

2.1.5 Jawny i ukryty FOA

U ludzi rozróżnia się dwa rodzaje kierunkowania uwagi: jawne (ang. *overt attention*) oraz ukryte (ang. *covert attention*). Różnica polega na sposobie jego działania [27].

Podczas „jawnego” kierunkowania uwagi pojawia się ruch gałki ocznej. Ponieważ wzrok człowieka widzi najostrzej na wprost, to co znajduje się dokładnie na linii widzenia jest niejako automatycznie w polu uwagi. Przykładowy rysunek przedstawiający obraz rzeczywisty oraz jaki dociera do mózgu przedstawiono na rys. 2.4.

Jawne kierunkowanie uwagi jest dzięki temu obserwowalne na zewnątrz – aby stwierdzić na co patrzy się człowiek wystarczy sprawdzić, w którą stronę ma skierowany wzrok. Własność tę wykorzystuje się podczas eksperymentów

(a) (b)

Rysunek 2.4: Przykład obrazu wejściowego (a) oraz odpowiadającego mu obrazu jaki dociera do mózgu (b) [27].

psychologicznych, mających na celu stwierdzenie na co i w jakiej kolejności patrzył się badany osobnik [45]. W szczególności tworzone są systemy wizyjne śledzące wzrok badanej osoby, potrafiące na tej podstawie określić, w które miejsce się w ona danej chwili patrzy z dokładnością lepszą niż 1°.

Podczas „ukrytego" kierunkowania uwagi nie następuje żaden ruch [27]. Z tego właśnie powodu ten rodzaj koncentracji nie jest łatwo mierzalny. Ze względu jednak na niską jakość obrazu poza głównym polem uwagi („jawna" uwaga – patrz rys. 2.4) jej znacznie jest drugoplanowe. Pozwala ona tylko na zauważanie bardzo istotnych lub nagłych zmian.

Połączenie uwagi jawnej i ukrytej jest często naśladowane poprzez jednoczesne stosowanie kamery szerokokątnej i silnie kierunkowej [41]. Więcej informacji na ten temat można znaleźć w sekc. 3.5.

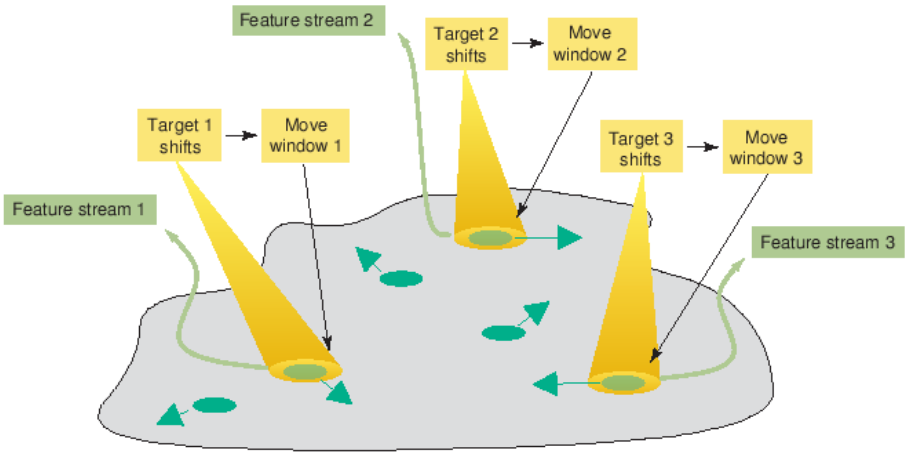
2.1.6 Śledzenie obiektów

Wykonano wiele ciekawych eksperymentów wymagających śledzenia obiektów przez człowieka. Choć w większości modeli zakłada się kierunkowanie uwagi na pojedynczym elemencie (np: [25]), podczas badań zaobserwowano, iż przeciętna osoba potrafi śledzić od 4 do 5 niezależnych obiektów przez kilkanaście sekund [39].

Możliwość śledzenia wielu obiektów jednocześnie jest bardzo porządaną cechą systemu wizyjnego a więc także i FOA. Zjawisko to jest nazywane ogólnie śledzeniem wielu obiektów (własne tłumaczenie ang. *multifocal attention*). Istnieje kilka koncepcji do jego realizacji, w tym [39]:

- pliki/teczki obiektów (własne tłumacznie ang. *object files*)
- grupowanie obiektów (ang. *grouping*) i śledzenie ich jako jednej całości
- przełączanie uwagi pomiędzy śledzone obiekty (ang. *shifting*)

W sposób ogólny zagadnienie może zostać zilustrowane jako kilka osobnych strumieni przypisanych do „okien" śledzących poszczególne obiekty w sposób od siebie niezależny (rys. 2.5).

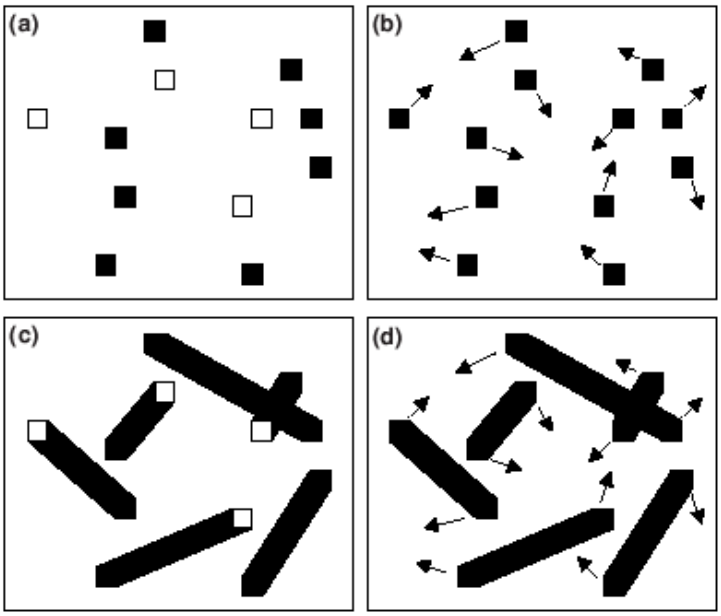


Rysunek 2.5: Ogólny model śledzenia wielu celów (obiektów) jednocześnie – szczegółowy opis zawarty w tekście (zaczerpnięte z [8]).

Ciekawą obserwację poczyniono w [11]. Autorzy pokazują, iż możliwości kierunkowania uwagi są wyuczalne i można je rozwijać. Dobrym przykładem są tu gry komputerowe. Badania pokazały większe możliwości śledzenia i mniejsze odległości pomiędzy obiektami w jakich są one nadal postrzegane jako osobne u graczy. Okazało się także, że juz po około 10 godzinach grania w gry akcji (wykorzystano do tego gry zręcznościowe, w których świat jest widziany z perspektywy pierwszej osoby – ang. *FPP – First Person Perspective*) widać wyraźny przyrost umiejętności percepcyjnych osoby grającej.

2.1.7 Grupowanie obiektów

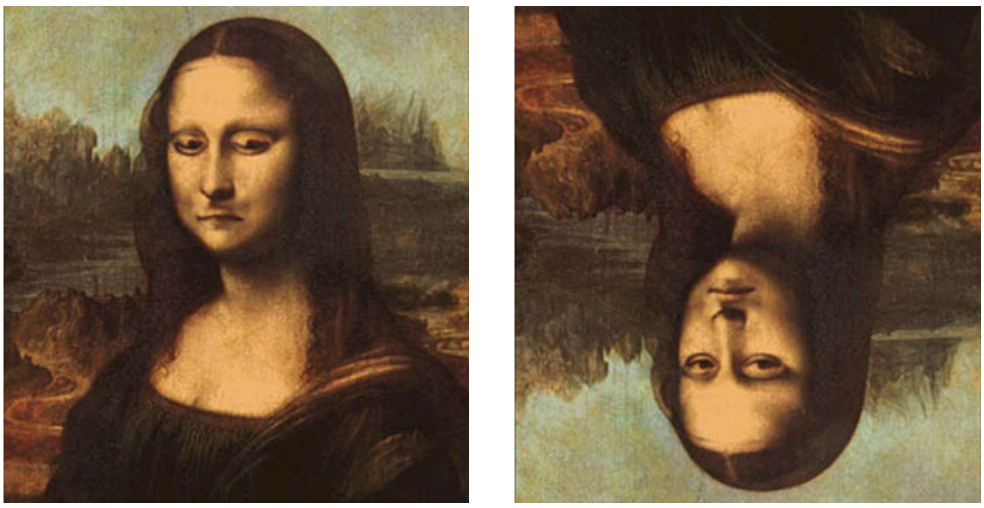
Eksperyment prezentowany w [8] pokazał, że podstawowym elementem śledzonym przez człowieka jest „obiekt" (co jest obiektem zależy od kontekstu). O ile śledzenie pojedynczych punktów nie sprawiało badanym trudności (były one traktowane jako osobne obiekty), połączenie ich liniami powodowało wymuszenie śledzenia połączonych elementów (obiektów) jako całości nie zaś ich pojedynczych punktów składowych. W praktyce powodowało to konieczność obserwowania rotacji i zmian kształtu figur nie pozwalając skupić się na pojedynczych punktach czyniąc zadanie podobne do poprzedniego znacznie trudniejszym. Przykład takiego eksperymentu pokazano na rys. 2.6.



Rysunek 2.6: Eksperyment ze śledzeniem obiektów zaproponowany w [8]: (a) warunki początkowe – jasne punkty mają być śledzone przez badanego (b) początek eksperymentu – punkty przemieszczają się w losowych kierunkach przez kilkanaście sekund po czym badany ma wskazać punkty które miał za zadanie śledzić (c) ten sam eksperyment, lecz punkty są połączone liniami (d) początek eksperymentu – przesówanie punktów powoduje zmianę kształtu i rotację figury utrudniając śledzenie (zaczerpnięte z [8]).

2.1.8 Zależność od kąta patrzenia

Powszechnie uważane jest, iż ludzki mózg jest w stanie rozpoznać obiekty niezależnie od kąta pod jakim są prezentowane [27][45]. Nie jest to zawsze parwdą. Ludzie przeważnie postrzegają świat w ten sam sposób (trawa na dole, chmury na niebie, etc…) co może utrudnić w pewnych sytuacjach rozpoznanie pewnych elementów. W szczególności obrót zdjęcia lub obrazu ludzkiej twarzy powoduje „wyłącznie" podświadomego mechanizmu, który wykrywa twarze na widzianych przez nas obrazach i analizuje je ze uwzględnieniem większej ilości detali (rozpoznawanie osób, ich nastrojów czy intencji). Powoduje to, iż ten sam obraz twarzy prezentowany pod różnymi kątami będzie postrzegany inaczej przez obserwatora. Przykład takiego zjawiska zaprezentowano na podstawie „koślawego" obrazu Monalizy, oraz jego pionowego odbicia rys. 2.7. Mimo, iż obrazy są binarnie identyczne twarz widziana odwrotnie jest postrzegana jako ładna, widziana „normalnie" zaś jako zniekształcona.



(a) (b)

Rysunek 2.7: Efekt różnicowania percepcji zależnie od kąta patrzenia: (a) obraz „normalny" (b) obraz odwrócony – mimo iż obrazy są binarnie identyczne, są postrzegane diametralnie inaczej przez ludzi.

2.2 Sieci neuronowe

Istnieje wiele pracy poświęconych kierunkowaniu uwagi bazujących na sieciach neuronowych [31][22][6][26]. Niemal wszystkie używają sieci WTA

(ang. *Winner-Take-All*) do wyznaczania miejsca najbardziej wyróżniającego się (ang. *the most salient*), jako miejsca skupienia uwagi [27][45][21].

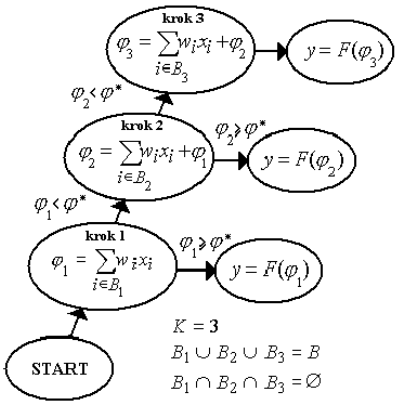
W niniejszym rozdziale zostanie pokrótce przybliżone działanie sieci WTA oraz zostanie (bardzo skrótowo) przedstawionych kilka przykładowych rozwiązań bazujących (niemal) wyłącznie na sieciach neuronowych[2].

2.2.1 Sieci dedykowane

W procesie wieloletnich badań powstało kilka modyfikacji modelu neuronów oraz całych sieci tak aby spełniały wymogi kierunkowania uwagi. W niniejszym opracowaniu zostaną skrótowo omówione dwa spośród nich [31][38].

Pierwszym z omówionych sieci będzie sieć sigma-if [31]. Jest to pewna modyfikacja perceptronu wielowarstwowego (ang. *Feed forward*) polegająca na zamianie „klasycznych" neuronów neuronami sigma-if, charakteryzującymi się zdolnością do kierunkowania uwagi.

Neuron taki posiada, podobnie jak perceptron prosty, wektor wag $w = [w_1, \ldots, w_n]$ oraz dodatkowo wektor grupujący $\theta = [\theta_1, \ldots, \theta_n]$ używany do wyliczania wartości funkcji agregacji. Na podstawie wektora θ ustalana jest kolejność sumowania ważonych (za pomocą wektora w) sygnałów wejściowych. Jeżeli suma przekroczy zadany próg ϕ^* następuje aktywacja neuronu. W przeciwnym przypadku liczona jest suma ważona dla kolejnej grupy, itd…Przykład takich obliczeń prezentuje rys. 2.8.



Rysunek 2.8: Przykład wyliczania wartości pobudzenia dla neuronu sigma-if (zaczerpnięte z [31]).

[2]Choć tematyka jest rozległa, ze względu na charakter tej pracy temat nie będzie dogłębnie rozwijany. W poszukiwaniu większej ilości informacji autor odsyła do bogatej literatury.

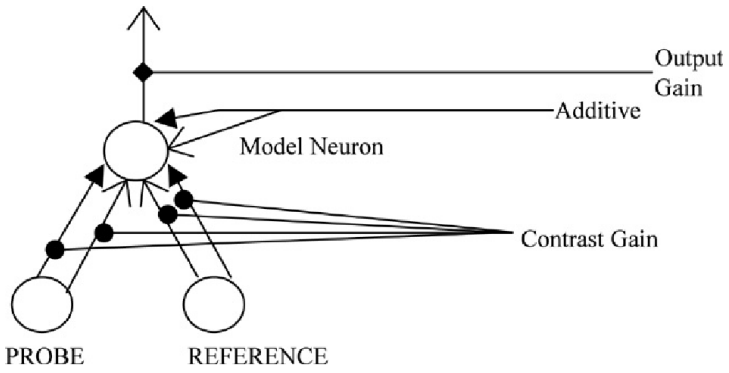
Sumarycznie daje to efekt kierunkowania uwagi ponieważ pewne sygnały mają pierwszeństwo przed innymi.

Uczenie takiej sieci wymaga modyfikowania wektorów wag w oraz grupującego θ na podstawie błędu dla przykładów uczących. Przykładami uczącymi są tu sekwencje kolejnych odwiedzanych punktów na danym obrazie. Całość operacji realizuje zaproponowany w [31] zmodyfikowany algorytm wstecznej propagacji błędu (ang. *Back Propagation*).

Drugie z prezentowanych podejść wymusza koncentrację uwagi w bardziej bezpośredni sposób – za pośrednictwem połączeń pobudzających oraz hamujących na różnych etapach przetwarzania informacji [38]:

- wejściach neuronu (ang. *constrast gain*)
- „wnętrzu" – bezpośrednie wpływanie na wartości wyliczane (ang. *additive*)
- wyjściu neuronu – modyfikacja wartości zwracanej (ang. *output gain*)

Przykład omawianego neuronu, wraz z zaznaczonymi pobudzeniami zewnętrznymi przedstawiono na rys. 2.9.



Rysunek 2.9: Neuron kierunkujący uwagę zaproponowany w [38] (zaczerpnięte z [38]).

Prezentowana sieć ma raczje charakter badawczy niż praktyczny. Z przeprowadzonych eksperymentów wynika, iż najlepsze efekty (ze względu na kierunkowanie uwagi) daje wpływanie poprzez modyfikowanie sygnału wejściowego (ang. *constrast gain*). Jest to również sposób kierunkowania uwagi z powodzeniem używany w [31].

2.2.2 Sieci WTA oraz IOR w FOA

Określenie „sieć WTA" (ang. *Winner Take All network*) nie jest jednoznaczne. Jest to skrót myślowy od sieci neuronowej, w której podczas pracy wyłaniany jest jeden „zwycięski" neuron.

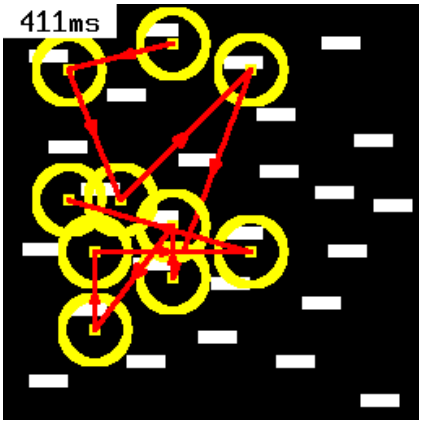
Istnieje kilka różnych modeli sieci spełniających paradygmat WTA. Jedną z nich jest sieć SOM (ang. *Self Organizing Map*), zwana także siecią Kohonena. Jest to sieć składająca się z neuronów posiadających przypisany wektor wag wejściowych oraz pozycję wewnątrz sieci. Typowo stosuje się sieci $2D$ – odpowiedź z takiej sieci może być wtedy interpretowana jako obraz rastrowy, gdzie każdemu punktowi na obrazie odpowiada pojedynczy neuron.

Właśnie taki sposób przetwarzania jest stosowany przy kierunkowaniu uwagi. Wejściowa mapa istotności jest podawana na wejście sieci na wyjściu zaś pobierana jest informacja o odpowiedzi – zwycięski neuron odpowiadający punktowi najbardziej istotnemu (najbardziej wyróżniającemu się). Punkt ten jest interpretowany jako miejsce do skierowania „uwagi" systemu.

Gdyby uruchomić taką siec „bezpośrednio" na mapie występności okazałoby się, iż pokazywany jest tylko jeden punkt przyciągający uwagę – zawsze ten sam. Jest to zjawisko niepożądane, ze względu na konieczność przeglądania również mniej „wyróżniających" się miejsc. W typowych sytuacjach interesujących jest kilka takich punktów, przeglądanych poczynając od najbardziej istotnych [45][27].

Aby uniemożliwić sieci ciągłe „wskazywanie" tych samych lokacji stosuje się mechanizm zabraniania powrotów (ang. *IOR - Inhibition Of Return*). W praktyce jest to realizowane poprzez wymuszenie zerowania odpowiedzi sieci w danym obszarze. Zależnie od implementacji, można to robić poprzez wyzerowanie obszaru w ustalonym, z góry zadanym promieniu od zwycięskiego neuronu [27] lub też poprzez resetowanie neuronów sąsiadujących ze zwycięskim jeśli ich odpowiedź jest silniejsza od pewnego zadanago a'priori progu (procedurę należy powtarzać rekurencyjnie dla wszystkich sąsiadów). Uzyskuje się w ten sposób efekt przełączenia uwagi na kolejny „istotny" fragment obrazu. Przykład przejścia FOA przez kilka kolejnych miejsc na obrazie przedstawiono na rys. 2.10.

Większość ważniejszych implementacji [45][27] stosuje IOR ograniczony czasowo w działaniu. Oznacza to, iż po krótkim czasie miejsce na obrazie jakie zostało „zabronione" ponownie może przyciągnąć uwagę. Jest to bardzo ważna własność z punktu widzenia przetwarzania sekwencji obrazów – jeśliby nie „zwalniać" zabronionych obszarów po pewnym czasie cały obraz byłby wyłączony z analizy i dalsza koncentracja uwagi nie byłaby możliwa.



Rysunek 2.10: Przykład przeskakiwania FOA po różnych elementach obrazu (zaczerpnięte z [2]).

2.3 Scieżki „where" oraz „what"

Podrozdział ten jest poświęcony przedstawieniu (w sposób poglądowy) ogólnej zasady działania zjawiska koncentrowania uwagi u ludzi. Z racji ukierunkowania tejże pracy na sztuczne systemy wizyjne, podane tu informacje mają charakter wzmiankowy. Autor zachęca czytelników zainteresowanych bardziej szczegółowymi opisami do zapoznania się z dostępną literaturą z dziedziny kierunkowania uwagi [28][27][45][34] oraz literatury medycznej.

Nasz mózg pobiera i przetwarza znaczące ilości informacji, głównie za pomocą narządu wzrokowego (ok. 90% [27]). Badania aktywności poszczególnych obszarów mózgu pozwoliły wyodrębnić obszary odpowiedzialne za określone funkcje.

Szkic mózgu z zaznaczonymi najważniejszymi kierunkami przesyłania informacji wizyjnej przedstawiono na rys. 2.11. Widać tam wyraźnie dwie ścieżki przesyłania informacji:

- A: wzrok $\rightarrow$V1,V2 $\rightarrow$MT $\rightarrow$PPC $\rightarrow$DLPFC
- B: wzrok $\rightarrow$V1,V2 $\rightarrow$V4 $\rightarrow$IT $\rightarrow$VLPFC

gdzie[3]:

- V1, V2, V4 (ang. *early cortical visual areas*) – pierwszorzędowe korowe ośrodki wzrokowe

[3]Autor pragnie w tym miejscu złożyć serdeczne podziękowania Pani neurolog Lidii Szurgot za nieocenioną pomoc w tłumaczeniu prezentowanej terminologii medycznej.

- MT (ang. *medial temporal area*) – pośrednie drogi przewodzące
- PPC (ang. *posterior parietal cortex*) – kora tylniej części płata ciemieniowego
- DLPFC (ang. *dorsolateral prefrontal cortex*) – obszar grzbietowo-boczny kory przedczołowej
- IT (ang. *inferotemporal cortex*) – drugorzędowe skroniowe korwe ośrodki wzrokowe
- VLPFC (ang. *ventrolaterial prefrontal cortex*) wewnętrzno-boczny obszar kory przedczołowej

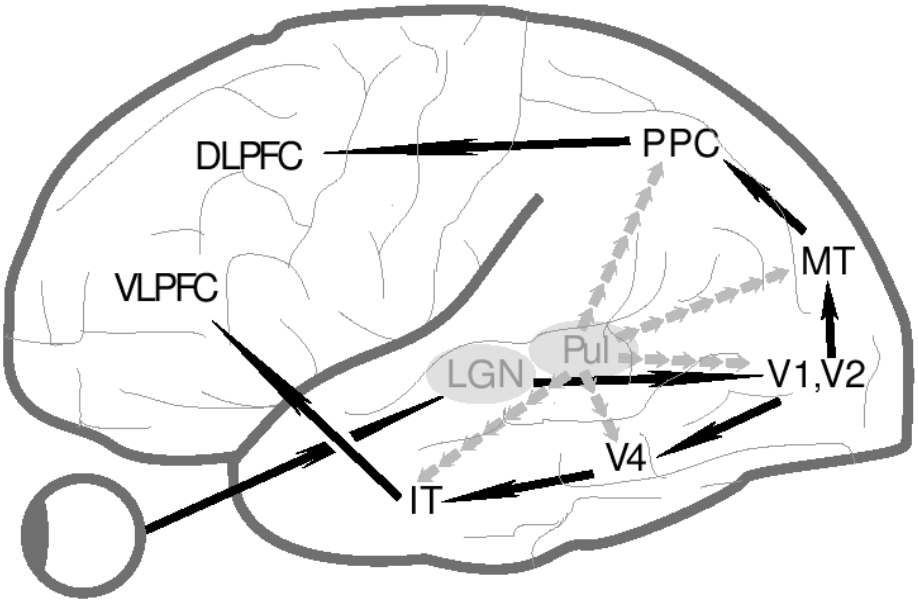
Badania pokazały, iż ścieżka A („górna" na prezentowanym szkicu mózgu) odpowiada za przekazywanie informacji o położeniu. Jest to więc ścieżka „where" („gdzie"). Dolna trasa (B) koduje informację o obserwowanym obiekcie – „what" („co").

Choć na rysunku ścieżki te nie łączą się ze sobą należy pamiętać, iż mózg nie zawiera idealnie „ostro" wydzielonych fragmentów nie połączonych z otoczeniem. W praktyce okazuje się, iż obie omawiane trasy sygnału wymieniają informacje między sobą. W szczególności oznacza to, iż odpowiedź pochodząca ze ścieżki „what" (np: rozpoznany obiekt) może zmodyfikować przebieg trasy kolejnych punktów odwiedzanych przez mechanizm kierunkowania uwagi [28].

Przykładem takiej sytuacji jest skrzyżowanie. Kiedy stoimy czekając na możliwość ruszenia (zielone światło) obserwujemy sygnalizator nad jezdnią. Kiedy światło zmieni kolor nasza uwaga jest ponownie skupiana na drodze mimo, iż nie ma ku temu przesłanek wynikających jedynie z widzianego przez nas obrazu (samych danych). Jest to typowy przykład zastosowania wiedzy semantycznej podczas sterowania uwagą.

Po zaprezentowaniu biologicznego przykładu systemu kierunkującego uwagę warto zadać sobie pytanie jak problem ten jest rozwiązywany przez maszyny.

W rzeczywistości sztuczne systemy wizyjne działają w ten sam sposób co ludzki mózg. W szczególności, na przykładzie systemów kierunkowania uwagi, znaleziony na obrazie punkt skupienia uwagi zawiera informację przestrzenną. Można to zinterpretować jako odpowiednik ścieżki „where". Z drugiej strony, po znalezieniu na obrazie danego fragmentu, jest on poddawany procesowi rozpoznawania przez jakiś klasyfikator. Ponieważ nie ma tu już jawnej informacji o położeniu w przestrzeni ale jest informacja dotycząca obserwowanego obiektu (efekt rozpoznania) można interpretować tę część systemu jako odpowiednik ścieżki „what".



Rysunek 2.11: Poglądowa ilustracja przepływu sygnału optycznego w mózgu człowieka. Dokładny opis zawarto w tekście (zaczerpnięto z [28]).

Rozdział 3

Stosowane podejścia

Rozdział ten opisuje najczęściej spotykane w zagadnieniach kierunkowania uwagi podejścia i stosowane przenośnie obrazujące niektóre z zagadnień (metafory). Poruszana jest także problematyka zastosowania wiedzy dziedzinowej do kierunkowania działania modeli jak i efekty jakie da się uzyskać bez niej. Prezentowane są także wybrane, stosowane w praktyce rozwiązania sprzętowe oraz sprzętowo-programowe. Podsumowaniem sekcji będzie przedstawienie kilku przykładów praktycznych systemów z FOA, realizujących konkretne zadania.

3.1 Metafory

W bieżącym podrozdziale jest opisanych kilka najbardziej znanych metafor koncentracji uwagi spotykanych w literaturze [33][27][45][34]. Choć metafory przeważnie nie są implementowane bardzo wiele publikacji powołuje się na nie jako na podstawę teoretyczną lub też jako bazową ideę leżącą u podstaw omawianego podejścia.

Choć wymienione metafory są najczęściej spotykanymi, nie są to wszystkie dostępne (pominięto np: bramkowanie uwagi (ang. *attention gating*) odnoszące się bardziej do aspektów czasowych a niżeli przestrzennych, prezentowanych poniżej). Osoby zainteresowane autor odsyła do literatury.

3.1.1 Światło punktowe

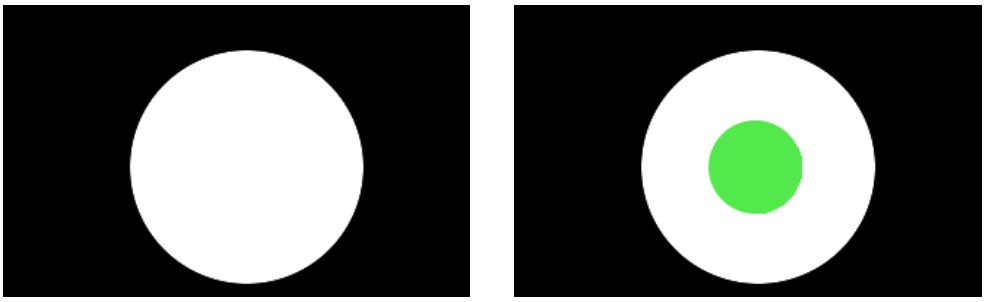
Jedną z pierwszych metafor ilustrujących działanie uwagi w systemach wizyjnych było „światło punktowe" (ang. *spotlight*). Metafora ta została zaproponowała w 1980 r. w [30]. Zakładała ona istnienie pojedynczego punktu kierunkowania uwagi, o stałym rozmiarze (koło o pewnej zadanej średnicy). Jest ona zgodna z podstawowymi eksperymentami, w których badani są stymulowaniu jednym

bodźcem wizualnym (np: silnym błyskiem światła) [5]. Jeśli rozmiar „interesującego” obszaru jest zgodny z założonym rozmiarem „światła”, zaznaczony w ten sposób obszar jest miejscem koncentrowania uwagi wizyjnej (patrz rys. 3.1).

(a) (b)

Rysunek 3.1: Przykładowa sytuacja kiedy metafora światła punktowego sprawdza się dobrze (rozmiar celu i założonego pola uwagi są zbliżone).

Niestety założenie stałości rozmiaru pola uwagi często nie jest spełnione (patrz rys. 3.2). W takiej sytuacji konieczne jest użycie innej metafory.



(a) (b)

Rysunek 3.2: Przykładowa sytuacja kiedy metafora światła punktowego nie sprawdza się (wyraźna różnica kształtu i rozmiaru celu i założonego pola uwagi).

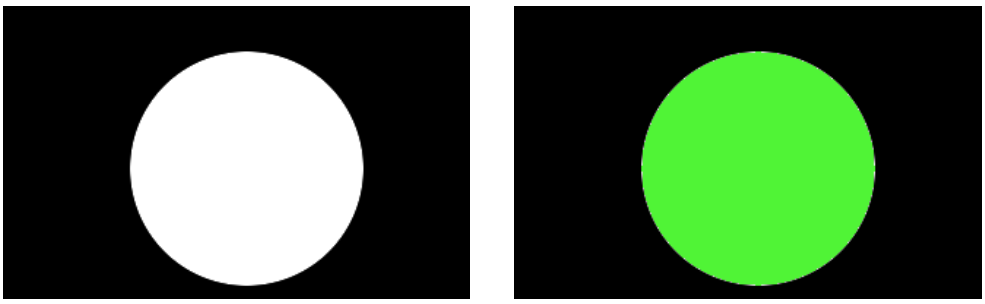
Jak słusznie zauważono w [33], choć metafora „światła punktowego” opisuje poprawnie tylko pewien niewielki zestaw eksperymentów, jest ona bardzo intuicyjna i chętnie stosowana jako wyjaśnienie pewnych sytuacji.

Przykładowym zastosowaniem jej jest ilustracja danych eksperymentalnych, gdzie zaznaczany jest punkt środkowy oraz jego otoczenie o pewnym stałym, założonym promieniu. Taka notacja jest stosowana między innymi w [27].

3.1.2 Soczewki powiększające

Podstawowy problem związany z metaforą światła punktowego (podsekc. 3.1.1) – założony z góry rozmiar obszaru koncentrowania uwagi – jest stosunkowo łatwy do obejścia. Należy umożliwić stosowanie zmiennego rozmiaru pola uwagi [12].

Właśnie ta obserwacja jest ideą metafory soczewek powiększających (ang. *zoom lens*). Sprawdza się ona dobrze, nadal pozostając bardzo intuicyjną i chętnie używaną do celów ilustracji [45][27]. Przykład jej funkcjonowania zaprezentowano na rys. 3.3. Co prawda metafora soczewek powiększających



(a) (b)

Rysunek 3.3: Przykładowa sytuacja kiedy metafora soczewek powiększających sprawdza się dobrze (pokrywa cały obszar).

jest znacznie bardziej elastyczna od światła punktowego, nadal nie rozwiązuje ona wszystkich możliwych sytuacji – nie jest „odporna” na kształty nie zbliżone do okręgu. Przykładową sytuację kiedy prezentowana metafora nie sprawdza się przedstawia rys. 3.4.

3.1.3 Gradient

Ponieważ metafory światła punktowego i soczewek powiększających nie są w stanie pokryć wszystkich możliwych sytuacji, na podstawie eksperymentów z obserwacją kierunkowania uwagi na literach i słowach zaproponowano w [15] metaforę gradientu. Podczas eksperymentów z udziałem ochotników zaobserwowano charakterystyczny sposób kierunkowania uwagi – jeśli mieli się oni skupiać na pojedynczych literach, obserwowano bardzo duży „pik” koncentracji na samej literze, który dość szybko zanikał wraz ze wzrostem odległości od centrum zainteresowania. W przypadku obserwowania całego wyrazu, gradient uwagi nie miał charakteru mocno punktowego, lecz był nieco łagodniejszy obejmując cały wyraz oraz, nieco szersze niż poprzednio, otoczenie.

(a) (b)

Rysunek 3.4: Przykładowa sytuacja kiedy metafora soczewek powiększających nie sprawdza się (kształt obiektu nie odpowiada kształtowi „soczewki”).

Metafor gradientu była więc kolejnym uogólnieniem – nie czyniła założeń co do kształtu obiektu przyciągającego uwagę ani do „ostrości” krawędzi obszaru zainteresowań (granica jest płynna).

3.1.4 Filtry VAP

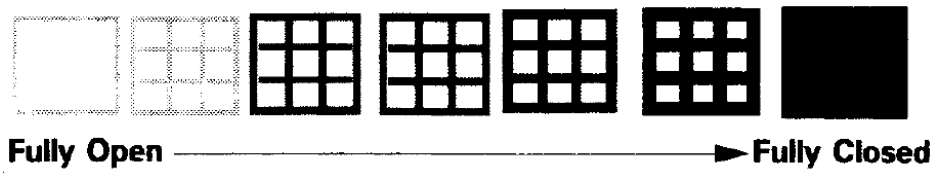
Choć idea metafory gradientu wydawała się spójna ze znaczną częścią przeprowadzanych eksperymentów nadal pozostawał jeden nie pokryty przypadek – uwaga dzielona pomiędzy kilka punktów na obrazie. Uwaga ta, zwana dalej uwagą wielomodlaną (własne tłumaczenie ang. *multifocal attention*), przejawia się w wielu różnych eksperymentach [4][39]. Pokazano, iż przeciętny człowiek jest w stanie śledzić około 4 niezależnych obiektów jednocześnie [39].

Z tego powodu w 1997 r. zaproponowano podejście zwane filtrami VAP (ang. *VAP filters*) [33], czyli przepuszczalnymi filtrami zmiennymi (ang. *Variable and Permeable Filters*. Zgodnie z tą koncepcją pole widzenia składa się z elementów dowolnego kształtu, dowolnego rozmiaru oraz dowolnej przepuszczalności informacji o danym obszarze. W sposób oczywisty jest to uogólnienie wszystkich prezentowanych w tym opracowaniu metafor [33].

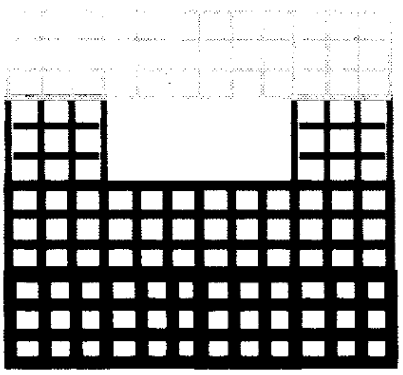
Przykład kilku filtrów o różnych stopniach przepuszczalności oraz różnym kształcie zaprezentowano kolejno na rys. 3.5 oraz rys. 3.6.

3.2 Mapy występności

Model mapy występności/istotności (ang. *saliency map*) jest podstawowym dla większości prac poświęconych kierunkowaniu uwagi [45][27][34]. W niniej-



Rysunek 3.5: Przykład serii filtrów VAP przepuszczających różną ilość informacji (zaczerpnięto z [33]).



Rysunek 3.6: Przykład serii filtrów VAP przepuszczających różną ilość informacji w różnych częściach obrazu (zaczerpnięto z [33]).

szym podrozdziale zostanie on bardziej szczegółowo opisany wraz ze spotykanymi w literaturze wariantami oraz rozszerzeniami.

3.2.1 Podstawy biologiczne

Badania nad wczesnymi etapami przetwarzania informacji wizyjnej pokazały ciekawą cechę ludzkiego mózgu. Zaobserwowano bowiem wrażliwość uwagi u ludzi na pewne szczególne cechy podstawowe (patrz rozdz. 2.1.1), które powodują, iż pewne elementy na scenie zauważamy od razu innych zaś musimy szukać przez dłuższą chwilę. Wyodrębniono 3 podstawowe elementy, które „zauważamy" i zwracamy na nie uwagę w sposób nieświadomy:

- kolory (konkretnie: czerwony, zielony, niebieski i żółty)[1]
- intensywność (jasność)
- orientacja (kąt)

Okazało się, że mózg ludzki zawiera obszary, które odpowiadają za wyszczególnianie tychże cech na widzianym obrazie – tworzone są odpowiadające im mapy cech (ang. *feature maps*). Na podstawie takich danych wyznaczany jest punkt najbardziej wyróżniający się i na niego jest kierowana uwaga.

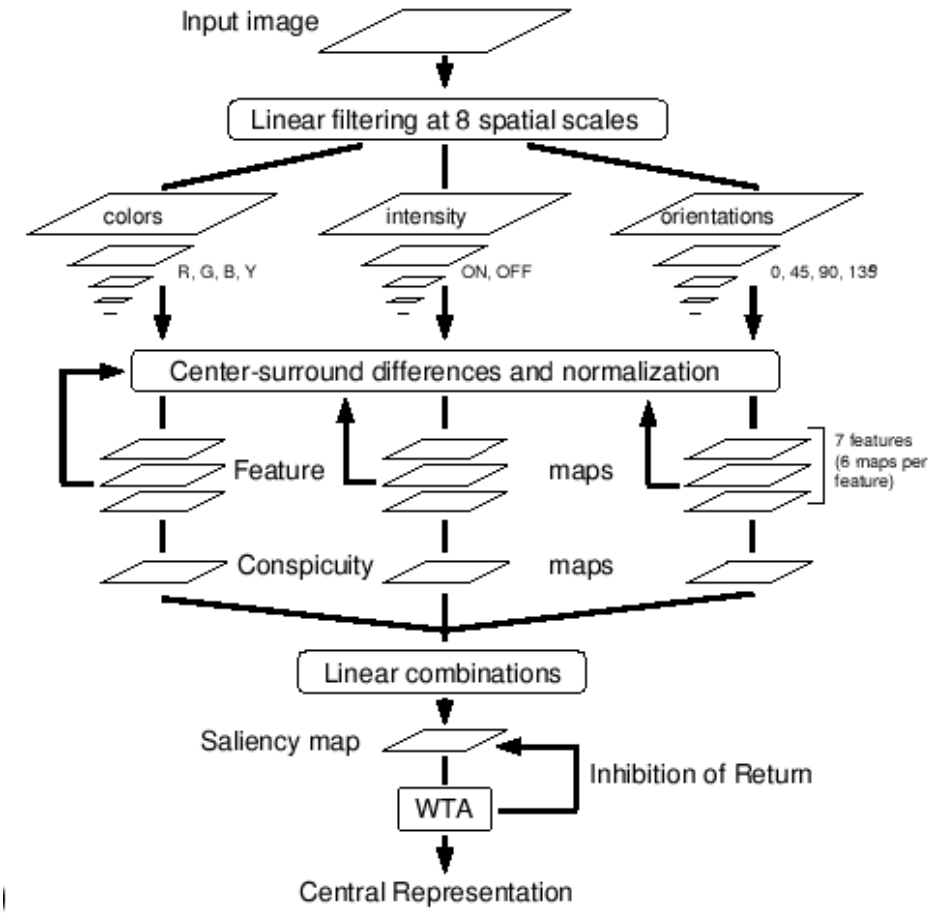
Obserwacja ta była inspiracją do stworzenia ogólnego, uniwersalnego modelu obliczeniowego.

3.2.2 Model obliczeniowy

Model mapy występności (ang. *saliency map* – tłumaczone także jako „mapy istotności" [31]) został zaproponowany w 1985 r. przez Koch'a i Ullman'a w [10]. Jego autorzy, wzorując się na naturze, stworzyli model obliczeniowy do wyznaczania miejsca koncentracji uwagi sterowany danymi (ang. *data driven; bottom-up*). Model ten jest przedstawiony na rys. 3.7.

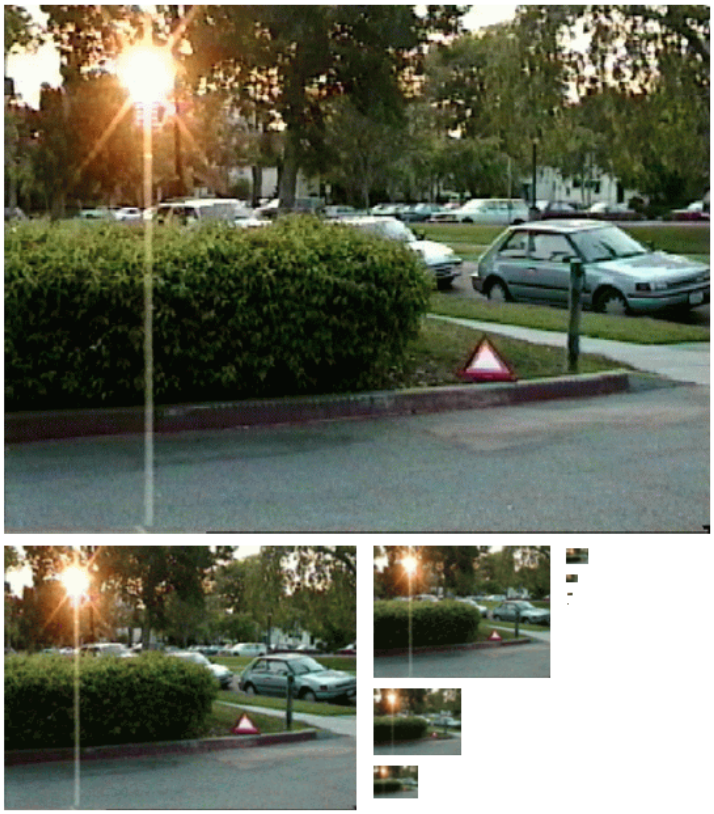
Ciekawy pomysł został wykorzystany do stworzenia map poszczególnych cech podstawowych. Pojedynczy obraz został kilkukrotnie zmniejszonych tworząc serię obrazów różnych rozmiarów. Wynik został zapisany w postaci tak zwanej piramidy obrazów (ang. *image pyramid*). Każdy z poziomów jest tworzony na podstawie poprzedniego po uprzednim przeskalowaniu go o zadaną skalę k (typowo przyjmuje się $k = 2$ lecz nie jest to wymagane).

[1] Należy w tym miejscu zauważyć, iż nie wszystkie kolory są równie istotne (wyróżniające się) dla człowieka. Np: znacznie szybciej reagujemy na kolor czerwony niż na zielony, itp [31][27]…



Rysunek 3.7: Mapa występności – opis w tekście (zaczerpnięte z [27]).

Samo skalowanie odbywa się w zasadzie za pomocą dowolnego algorytmu, jednak przy jego doborze należy mieć na uwadze zjawiska towarzyszące takie jak utrata informacji i dokładności. W szczególności niedostateczne próbkowanie (ang. *under sampling*) może spowodować istotną modyfikację obrazu oryginalnego [45]. Jeśli przeskalować np: czarnobiałą szachownicę poprzez branie do obrazu wyjściowego co drugiego punktu ($k = 2$) uzyskany obraz byłby biały lub czarny (zależnie od szczegółów implementacji algorytmu) co znacząco zmieniłoby jego sens. Efekt ten nazywa się potocznie aliasingiem (ang. *aliasing*). Aby uniknąć tego typu efektów stosuje się algorytmy skalujące wyliczające docelowy punkt na podstawie kilku punktów obrazu wejściowego, lub też przed samym skalowaniem przepuszcza się obraz przez filtr uśredniający. W ten sposób są tworzone najczęściej spotykane w praktyce piramidy Gauss'a (ang. *Gaussian pyramid*) oraz Laplace'a (ang. *Laplacian pyramid*) [27][45]. Przykład piramidy obrazów dla obrazu rzeczywistego oraz $k = 2$ ilustruje rys. 3.8



Rysunek 3.8: Przykładowa piramida obrazów dla $k = 2$ (zaczerpnięte z [27]).

Piramidy obrazów [47] stosuje się powszechnie jako alternatywę dla znacznie bardziej kosztownych obliczeniowo serii dużych masek. Aby wydobyć z obrazu fragmenty na różnym stopniu szczegółowości, zamiast przetwarzać go kilkukrotnie za pomocą filtrów o coraz to większym rozmiarze, skaluje się obraz wejściowy kilkukrotnie tworząc piramidę. Następnie dla każdego poziomu piramidy stosuje się dokładnie ten sam filtr.

Mając wyliczone mapy dla wszystkich podstawowych cech (ang. *feature maps*) w różnych skalach należy je znormalizować do wspólnej wielkości, aby jeden obraz zawierał cechy zarówno z obrazów z niską ilością detali (mała rozdzielczość) jak i wysoką (duża rozdzielczość). Typowo w literaturze spotyka się kilka podejść do tego problemu – wszystkie obrazy są skalowane do rozmiaru najmniejszego z nich, największego lub środkowego [27][45]. po wykonaniu takich operacji obrazy można dodać piksel po pikselu. W ten sposób uzyskujemy pojedynczy obraz dla pojedynczej cechy (np: jasności), zwany mapą widoczności (ang. *conspiciuty map*).

Kolejnym etapem jest unormowanie wszystkich map widoczności do tego samego zakresu wartości (np: $0 - 1$), tak aby można je było ze sobą połączyć. Operacja ta nazywa się normalizacją – oznaczmy ją jako $N(\cdot)$. Istnieje wiele sposobów na wykonanie tej operacji. Najważniejsze spośród nich zostały omówione w podsekc. 3.2.3.

Mając pojedynczy obraz końcowy, zwany mapą występności (ang. *saliency map*) można przejść do końcowego etapu – wyznaczania pojedynczego punktu uwagi (FOA). W oryginalnym modelu [10] oraz ciekawszych jego późniejszych implementacjach [27][45] stosuje się do tego celu sieć WTA wraz z mechanizmem zabraniania powrotu IOR (patrz podsekc. 2.2.2). Odpowiedzią z takiej sieci jest pojedynczy punkt, który jest traktowany jako miejsce kierunkowania uwagi.

3.2.3 Normalizacja map widoczności

Mając kilka map widoczności (ang. *conspiciuty maps*) należy dokonać ich fuzji celem stworzenia pojedynczej mapy (mapy występności), z której będzie można wprost odczytać miejsce kierunkowania uwagi. Operacja ta ma klasycznie za zadanie preferować mapy o określonej charakterystyce – powinna preferować mapy posiadające jeden (do kilku) wyraźnych pików nad mapy zawierające dużo podobnych maksimów lokalnych.

W litearturze spotyka się kilka różnych sposobów na rozwiązanie tego problemu. Poniżej zostaną omówione wybrane spośród nich.

Naiwne sumowanie (ang. *naive sumation*) został opisany w [27]. Polega on na prostym sumowaniu obrazów, które są unormowane do tego samego zakresu wartości (np: $0-1$). Przeważnie każda pojedyncza cecha zawiera wiele lokalnych minimów i relatywnie dużo szumu. Opisana tu metoda nie sprawdza się więc najlepiej. Jej stosowanie jest odradzane [27].

Uczenie kombinacji liniowej (ang. *learning linear combination*) to drugie z podejść prezentowanych w [27]. Jest to metoda nadzorowanego uczenia przypisującego wagi poszczególnym mapom widoczności. Jako dane wejściowe służą obrazy ze zdefiniowanymi przez użytkownika, odpowiadającymi punktami istotnymi (przyciągającymi uwagę).

Mapy są mnożone przez wyznaczony w ten sposób współczynnik i dopiero potem sumowane [27]. Daje to przewagę nad podejściem naiwnym, gdyż sumowanie zawiera już pewne współczynniki określające jakie elementy były ważniejsze w procesie uczenia. Niestety odbywa się to kosztem uniwersalności – operator $N(\cdot)$ jest przystosowywany do pewnej klasy zadań. Można powiedzieć, że zawiera w sobie wiedzę dziedzinową.

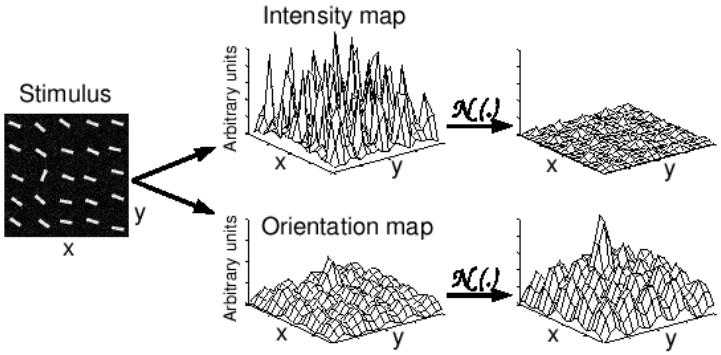
Choć efekty jego stosowania są o lepsze od naiwnego sumowania, wyniki są nadal odległe od obserwowanych u ludzi [27].

Globalne wzmacnianie nieliniowe na podstawie zawartości (ang. *Contents-based global non-linear amplification*) to kolejne z podejść proponowanych przez Itti [27]. Procedura dla tego operatora $N(\cdot)$ przebiega następująco:

1. dla każdej z map widoczności wyznaczyć globalne maksimum M oraz średnią z lokalnych maksimów $\bar{m}$
2. pomnożyć każdą z map przez wyrażenie $(M-\bar{m})^2$
3. zsumować wszystkie mapy

Przykład działania zaprezentowanej procedury pokazano na obrazie z którego wyliczono 2 mapy widoczności – obrazową dla danego przypadku (wyraźny pik) i nie niosącą informacji (dużo podobnych maksimów). Całość zilustrowano na rys. 3.9

Powyższa procedura ma kilka zasadniczych wad. Jak zauważono w [27][45] taka normalizacja nie będzie dawała oczekiwanych wyników jeśli na mapie będą dwa maksima o zbliżonej wartości (wyrażenie do przemnożenia mapy spowoduje ich wyzerowanie). Wymaga ono również wykonania pewnych globalnych



Rysunek 3.9: Przykłada działania normalizacji $N(\cdot)$: globalne wzmacnianie nieliniowe na podstawie zawartości (zaczerpnięte z [27]).

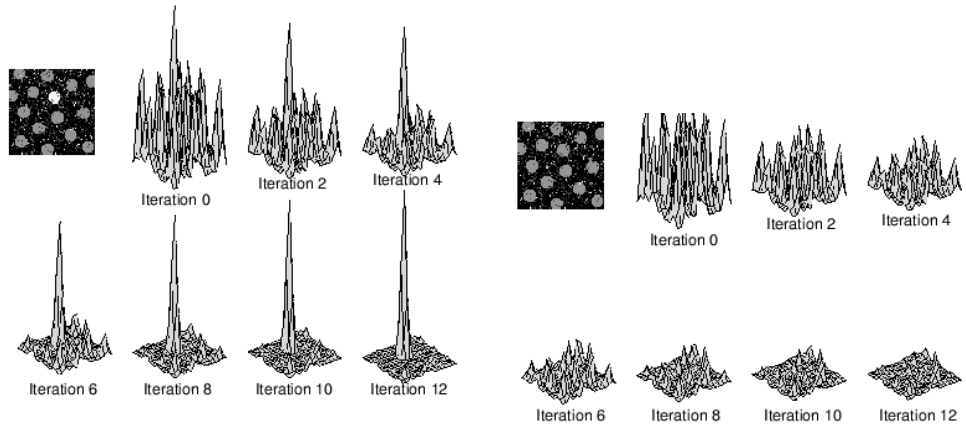
obliczeń na całej mapie (np: wyliczenie $\bar{m}$). Nie jest to problemem przy implementacji na komputerze sekwencyjnym, jednak myśląc o systemach wizyjnych w szerzej perspektywie należy mieć na uwadze ich specyficzny charakter – dużą ilość danych do przetworzenia w podobny sposób (np: poprzez filtry). Patrząc na to w ten sposób od razu wyłania się wizja komputerów o bardzo wysokim stopniu zrównoleglenia, umożliwiających przetwarzanie w czasie rzeczywistym. Niestety korzystanie z cech globalnych nie daje możliwości efektywnego zrównoleglenia[2], w związku z czym stanowi wąskie gardło przy takim podejściu.

Iteracyjne lokalne interakcje (ang. *Iterative localized interactions*) to czwarte i ostatnie z prezentowanych w [27] podejść. Procedura normalizacji polega na iteracyjnym przetwarzaniu obrazu za pomocą dwuwymiarowych filtrów DoG[3] (ang. *Difference of Gausians*). Efekt działania omawianego podejścia przedstawiono na rys. 3.10.

Metoda ta, choć kosztowna obliczeniowo (wielokrotne iterowanie po obrazie) jest jedyną zaprezentowaną w [27] dającą zadowalające rezultaty. Jedynym punktem dyskusyjnym jest ilość iteracji jaką należy wykonać. Choć znormalizowana mapa stabilizuje się juz po kilku iteracjach w przypadku istnienia wyraźnego maksimum (rys. 3.10(a)) wymaga nieco większej ilości

[2]Wyliczanie globalnego maksimum da się przyśpieszyć poprzez zrównoleglenie, uzyskując złożoność $O(\ln n)$ zamiast $O(n)$, jak dla przypadku sekwencyjnego, jednak dla dużego n jest to nadal sytuacja nieakceptowalna w porównaniu z filtrowaniem mającym w tym przypadku złożoność $O(1)$.

[3]Podejście takie zostało przyjęte ze względu na jego zgodność z charakterystyką percepcji ludzkiego oka. Jako przypadek wzorowany na naturze, nie będzie on dalej rozwijany. Osoby zainteresowane autor odsyła do lektury oryginału [27].



(a) (b)

Rysunek 3.10: Przykłady przetwarzania dwóch różnychobrazów za pomocą metody iteracyjnych lokalnych interakcji: (a) przykład z jednym wyraźnym maksimum; (b) brak pojedynczego globalnego maksimum (zaczerpnięte z [27]).

iteracji w przypadku jego braku (rys. 3.10(b)). Dobór tegoż parametru nie jest więc oczywisty.

Dzielenie przez ilość[4] zostało zaproponowane w [45]. Polega na zliczeniu sumy maksimów lokalnych na danej mapie widoczności, które przekraczają pewien założony próg m (w [45] zaproponowano ustalenie progu na poziomie 50% wartości największego z maksimów). Mając wyliczone m normalizujemy obraz X dla każdego punktu wyliczając:

$$W(X) = X/\sqrt{m}$$

Prezentowane podejście zachowuje się zgodnie z intuicją, promując mapy o małej ilości wyraźnych maksimów, jednocześnie nie powodując „dziwnych" efektów ubocznych jak prezentowana w [27] metoda globalnego wzmacniania nieliniowego na podstawie zawartości. Wadą tejże metody jest jednak jej globalny charakter – zaimplementowanie jej na komputerze wieloprocesorowym, dedykowanym do obróbki obrazów okaże się mieć złożoność co najmniej $O(\ln n)$.

Prezentowany tutaj przegląd metod normalizacji nie jest kompletny, nie mniej prezentuje najważniejsze podejścia. Ponieważ normalizacja jest jednym z istot-

[4]Proponowana nazwa na sposób normalizacji zaprezentowany (lecz nie nazwany) w [45].

niejszych elementów przetwarzania za pomocą modelu mapy występności, warto poświęcić czas na zaproponowanie nowych, skuteczniejszych operatorów.

3.2.4 Dalsze rozszerzenia

Klasycznie do budowy map istotności stosuje się, prezentowany w podsekc. 2.1.1 zestaw cech podstawowych. Nie jest to jednak jedyna możliwość. Od czasu zaprezentowania pionierskiej wersji modelu powstało wiele rozszerzeń i modyfikacji. Poniżej zostaną omówione najciekawsze z nich.

Głębia obrazu jest doskonałym przykładem dobrego źródła informacji, jakie można wykorzystać do kierunkowania uwagi. Posiada ona również uzasadnienie biologiczne – to co się znajduje bliżej, ma większą szansę bycia zauważonym, niż coś znajdującego się w większej odległości.

Istnieje kilka powszechnie spotykanych metod na odczytanie głębi z otoczenia. W szczególności można to robić za pomocą:

- stereowizji – na podstawie sygnału z 2 kamer istnieje możliwość odczytania informacji o odległości [40] [29].
- stereowizji trójkamerowej (3 kamery tworzące 2 pary: pionową oraz poziomą) – rozszerzenie systemu stereowizyjnego o 3 kamerę [34]. Rozwiązanie to jest przydatne w rozwiązywaniu sytuacji niejednoznacznych jakie mogą się pojawić w przypadku obrazu z 2 kamer.
- monowizji – choć teoretycznie nie jest to możliwe, w praktyce da się uzyskać przybliżony obraz głębii stosując pewne uproszczenia [9] bądź heurystyki [7].
- skanerów laserowych $3D$ [45] – metoda ta jest dokładniejsza od powyższych, jednak skanowanie trwa kilka sekund [45] co praktycznie uniemożliwia zastosowanie tej metody w zadaniach czasu rzeczywsitego.
- kamer $3D$ – rozwiązanie to jest uogólnieniem skanerów $3D$, będącym obecnie w fazie eksperymentalnej [45]

Inne modele barw niż najpopularniejszy RGB są również stosowane w celu poprawy wyników uzyskiwanych przez systemy FOA. Model RGB nie nadaje się dobrze do przetwarzania obrazów, ponieważ nie rozdziela on informacji chromatycznej od intensywności [45]. Rozdzielenie to posiada np: model HSV (ang. *Hue Saturation Value (intensity)*) oraz model LAB [45]. Ostani z prezentowanych jest szczególnie zalecany w [45] jako najlepiej odpowiadający charakterystyce ludzkiego oka.

LAB zawiera informacje o intensywności (L), pozycji pomiędzy czerwonym a zielonym (A) oraz pozycji pomiędzy zielonym a żółtym (B). Kolor zapisany w modelu LAB może być wyznaczony bezpośrednio z modeli XYZ, ten zaś z RGB [45].

Ruch jest niezwykle istotnym elementem pod kątem kierunkowania uwagi u ludzi. Dlatego też jest on coraz częściej uwzględniany w nowszych pracach poświęconych kierunkowaniu uwagi na obrazach [23][32][34].

Szczególnie ciekawe połącznie zaprezentowano w [34]. Do wyznaczania mapy istotości na podstawie ruchu wykorzystano bowiem jeden z algorytmów wyliczania pływu optycznego (ang. *optical flow* [13][44]). Zaproponowano także rozszerzenie modelu mapy występności poprzez dodanie aspektów związanych z ruchem (ang. *dynamic visual attention*).

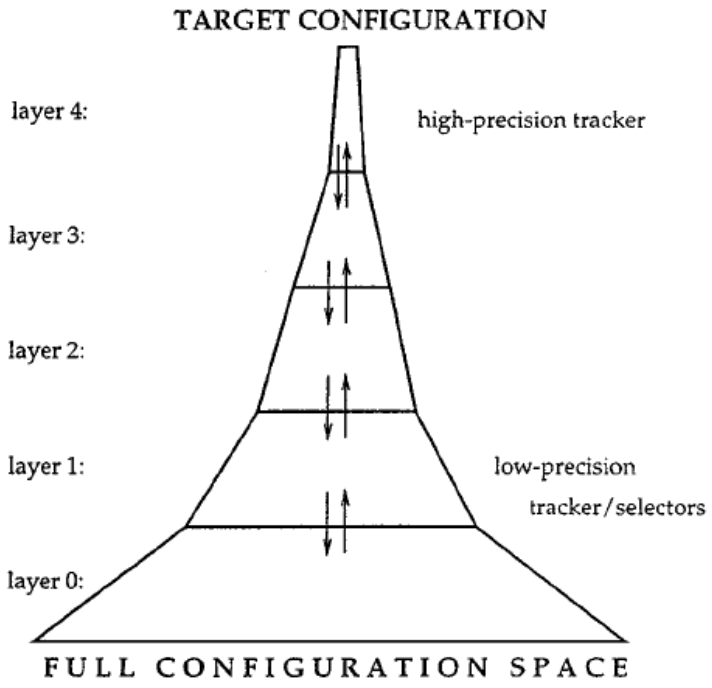
3.3 Model inkrementacyjny

Ciekawy model kierunkowania uwagi został przedstawiony w 1999 r. w [25]. Zakłada on posortowanie grupy „wybieraczy" (ang. *selectors*) oraz „podążaczy" (ang. *trackers*) począwszy od najbardziej ogólnych na najbardziej szczegółowych skończywszy. Zadaniem „wybieraczy" jest odnalezienie na obrazie interesujących fragmentów (np: głowy człowieka). „Podążacze" zaś mają za zadanie śledzić wybrany obiekt.

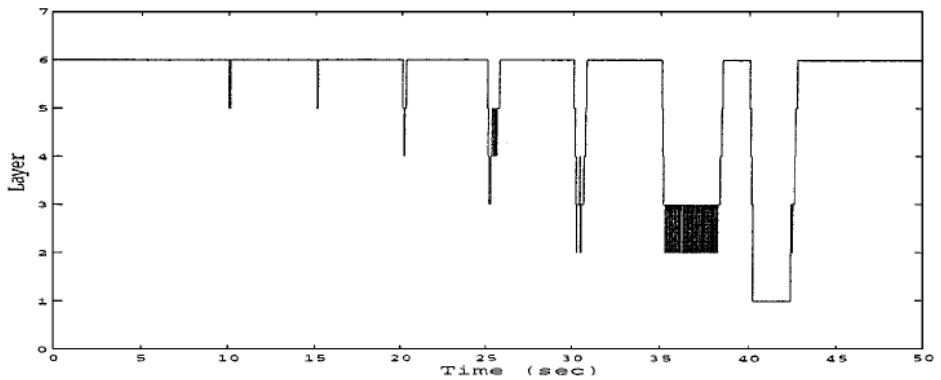
Praca modelu zaczyna się od uruchomienia najbardziej ogólnego spośród wybieraczy. Jeżeli uda mi się odnaleźć zadany obiekt jest on śledzony przez najbardziej ogólny z „podążaczy". W każdej kolejnej iteracji wybierany jest kolejny, bardziej szczegółowy „wybieracz" oraz „podążacz". Jeżeli bardziej szczegółowy „wybieracz" lub podążacz nie podoła swojemu zadaniu, następuje cofanie o jeden poziom wyżej w hierarchii – w kolejnej iteracji używany jest bardziej ogólny „wybieracz" lub „podążacz" (zależnie od tego, który z nich zawiódł).

Stos takich elementów („wybieraczy" lub „podążaczy") przedstawiono w sposób schematyczny na rys. 3.11.

Podczas działania w warunkach normalnych (tzn: dobrej „widoczności" szukanego i śledzonego obiektu) system pozostaje cały czas na najwyższym (tu: 6) poziomie pracy. Co $5[s]$ wprowadzane jest pewne zaburzenie takie jak gwałtowny ruch śledzonego przedmiotu lub jego (częściowe) zasłonięcie. Proces ten ilustruje rys. 3.12. Widać na nim wyraźnie, iż system reaguje na napotkany problem zmniejszeniem stopnia szczegółowości używanego „wybieracza"/„podążacza" aż do momentu odnalezienia szukanego elementu, kiedy to ponownie próbuje zastosować rozwiązanie bardziej szczegółowe. Po ustąpieniu zakłócenia system ponownie wraca na najwyższy poziom szczegółowości.



Rysunek 3.11: Przykład stosu elementów (tu 5 poziomowego), używanego w modelu (zaczerpnięto z [25]).



Rysunek 3.12: Przykładowy wykres poziomu pracy „podążacza” (z którego poziomu szczegółowości jest brany) w funkcji czasu. Co $5[s]$ wprowadzane jest jakieś zakłócenie – np: chwilowe zasłonięcie obiektu (zaczerpnięto z [25]).

3.4 FOA sterowane danymi i celami

Mówiąc o kierunkowaniu uwagi wszystkie istniejące podejścia można sprowadzić do dwóch przypadków:

1. sterowania kierunkowanego danym – „z dołu do góry" (ang. *bottom-up*)
2. sterowania kierunkowanego celami – „z góry do dołu" (ang. *top-down*)

Przypadek pierwszy, sterowany danymi, zakłada brak jakiejkolwiek wiedzy dziedzinowej odnośnie rozwiązywanego problemu. Jest to podejście bardzo ogólne, nie zależne od rozpatrywanego problemu.

Choć typowo stosowanie wiedzy dziedzinowej jest wymagane, istnieje pewna podgrupa problemów dla której dodatkowa wiedza nie jest konieczna. Przykładem należącym do tej grupy jest problem odnajdowania znaków drogowych na zdjęciach. Ponieważ znaki drogowe są z założenia projektowane z myślą o łatwości ich dostrzegania, nawet bez wiedzy dziedzinowej łatwo jest odnaleźć miejsca, w których mogą się one znajdować [27][45]. Niejako „wyłaniają" się one z otoczenia czyniąc je łatwymi do wykrycia. Oczywiście na kolejnych etapach konieczne jest użycie klasyfikatora aby zweryfikować hipotezę o występowaniu znaku w danym miejscu obrazu, jednak poszukiwanie ma bardzo zawężony charakter (ogranicza się do zaznaczonego obszaru).

Drugie z wyróżnionych podejść, kierunkowanie sterowane celami, jest koniecznością w większości systemów praktycznych, gdyż umożliwia nastawienie pracy systemu pod kątem wykonywania określonych czynności. Mając dostęp do metod kierunkowania na określone cechy, uzyskujemy elastyczny system do przetwarzania obrazu.

Przykładem takiego rozszerzenia w systemie VOCUS [45] są wagi uczone dla poszczególnych map widoczności, pod kątem eksponowania cech charakterystycznych danego, poszukiwanego elementu. Dzięki temu możliwe staje się kierunkowanie uwagi ze szczególnym naciskiem na wybrane cele (poprzez odpowiadający im zestaw predefiniowanych wag).

Oczywiście możliwe jest połączenie analizy sterowanej danymi (podstawa) oraz sterowanej celami (umożliwienie kierunkowania poszukiwań). Właśnie takie podejścia dają w praktyce najlepsze rezultaty [27][45].

Prezentowany w sekc. 3.3 model mapy istotności, w oryginalnej wersji [10] jest typowym przykładem podejścia sterowanego danymi. Posiada on jednak pewne rozszerzenia, które pozwalają na kierunkowanie poszukiwań i dopasowanie do konkretnego problemu [27][45].

Podobnie model przedstawiony w [25] (sekc. 3.3), który nie narzuca stosowania wiedzy dziedzinowej, choć w praktyce stosowanie konkretnych markerów (ang. *selectors*) oraz podsystemów śledzących (ang. *trackers*) przeważnie będzie wymagało zastosowania takowej (przynajmniej na poziomie bardziej szczegółowych metod).

3.5 Rozwiązania sprzętowo-programowe

Grafika rastrowa to macierz punktów reprezentujących właściwości wizyjne danego wycinka obserwowanej rzeczywistości. Choć sama reprezentacja punktów jest niewielka (typowo do kilku bajtów) jest ich bardzo dużo (obraz $800x600$ pikseli to 480000 punktów). Ponieważ znaczna część metod obróbki obrazów operuje na konkretnym punkcie oraz jego najbliższym otoczeniu, sensowym wydaje się przyśpieszanie tego typu obliczeń stosując specjalizowane platformy sprzętowe. Stosowanie tego typu podejść ma jeszcze jedną zaletę z punktu widzenia praktycznych systemów wizyjnych – niewielki pobór prądu względem sekwencyjnych maszyn cyfrowych ogólnego przeznaczenia.

Ostatnimi laty daje się zaobserwować rosnące zapotrzebowanie oraz zainteresowanie problematyką sprzętowego wspierania obliczeń systemów wizyjnych oraz neuronowych[5]. Problem implementacji sprzętowych oraz sprzętowo-programowych jest coraz częściej poruszany w literaturze [9][26][35][34]. Niedawno pojawiła się też polskojęzyczna książka dedykowana tejże problematyce [24]. Zjawisko to wydaję się być naturalnym, zważywszy na fakt, iż większość otoczającej nas na codzień elektroniki to urządzenia wbudowane. Wejście systemów wizyjnych w tą dziedzinę wydaje się być jedynie kwestią czasu. Już obecnie pojawiają się dedykowane do systemów wizyjnych układy [24][35] oraz systemy wizyjne pracujące na urządzeniach zasilanych bateryjnie [3].

W tym podrozdziale zamieszczono skrótowy opis kilku przykładów systemów wizyjnych z kierunkowaniem uwagi, wykorzystujących sprzętowe akceleracje obliczeń lub inne dedykowane rozwiązania.

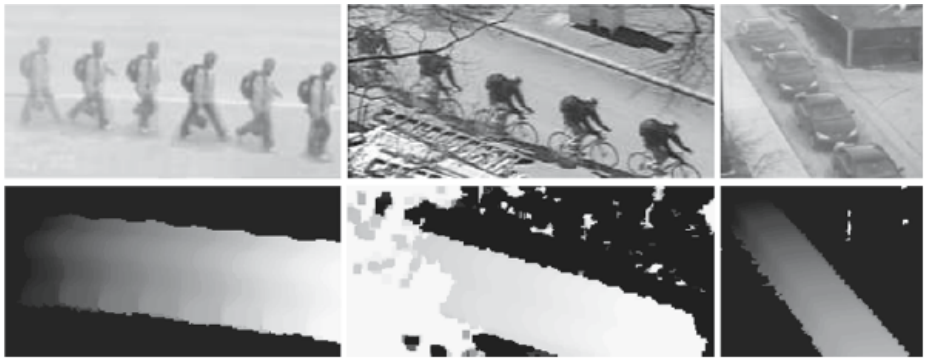
Najprostszym przykładem kooperacji systemu kierunkowania uwagi ze sprzętem jest sterowanie kamerą w celu śledzenia „zauważonych" proto-obiektów[6]. Prezentowany w [41] system składa się z 2 kamer: jednej szerokokątnej oraz

[5]Choć obliczenia w systemach wizyjnych za pomocą sieci neuronowych nie są jedynymi, są one niezwykle często spotykane [9][26].

[6]Proto-obiekty (ang. *proto-objects*) to elementy obrazu jakie zostały znalezione przez system kierunkowania uwagi. W pewnych sytuacjach mogą się one pokrywać z obiektami w znaczeniu znanym z problematyki rozpoznawania obiektów (ang. *object recognition*), jednak nie jest to prawdą w przypadku ogólnym [17].

jednej kierunkowej, dającej obraz wysokiej jakości ale o niewielkim promieniu widzenia. Łatwo zauważyć, że ów system jest podobny do działania ludzkiego oka – widzimy dużo dookoła, lecz w słabej jakości oraz bardzo wyraźnie dokładnie na wprost („jawne kierunkowanie uwagi").

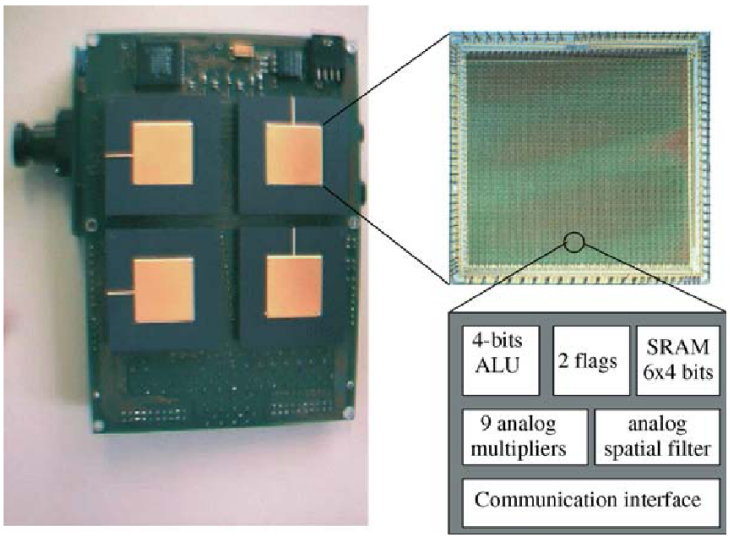
Innym prostym przykładem, kiedy FOA współpracuje ze sterowaniem kamerą jest system zaproponowany w [23]. W rozwiązaniu tym pokazano jak można poprawić efektywność systemów monitorujących poprzez zastosowanie kamery mogącej w sposób autonomiczny śledzić pewne proto-obiekty w rzeczywistym świecie. Efektem pracy przedstawionego w [23] systemu jest zaznaczenie tras poruszających się obiektów przy jednoczesnym wytłumianiu tła (obiektów nie poruszających się). Przykład pracy takiego systemu dla kilku scen pokazano na rys. 3.13



Rysunek 3.13: Przykład efektów pracy systemu do zaznaczania proto-obiektów w ruchu (zaczerpnięte z [23]).

Znacznie bardziej zaawansowane rozwiązanie zostało zaprezentowane w [35]. System wizyjny bazujący na modelu mapy występności (sekc. 3.2) został zaimplementowany na dedykowanej platformie sprzętowej o nazwie „ProtoEye" (rys. 3.14). „ProtoEye" jest architekturą typu SIMD (ang. *Single Instruction Multiple Data*), doskonale nadającą się do obliczeń na obrazach rastrowych.

Badania eksperymentalne systemu pokazały, iż system ten, mimo pracy z częstotliwością zaledwie $4[MHz]$ jest w stanie osiągnąć prędkość przetwarzania na poziomie 14 klatek na sekundę. Dodatkowo okazało się, iż głównym elementem opóźniającym jest transfer do i z pamięci, który odbywa się sekwencyjnie – jest więc wąskim gardłem całości. Z pomiarów wynika, iż same obliczenia zajmują zaledwie kilka procent całego czasu przetwarzania, zaś wymiana danych z pamięcią zewnętrzną zajmuje około połowy całości [34].



Rysunek 3.14: Zdjęcie systemu „ProtoEye” z przedstawioną schematycznie budową pojedynczego procesora obliczeniowego (zaczerpnięte z [35]).

System częściowo sprzętowy, częściowo programowy zaproponowano również w [26]. W tej implementacji zastosowano dedykowany układ do obliczeń na sieciach neuronowych, realizujących paradygmat sieci WTA. Pozostała część obliczeń – wyznaczanie mapy istotności – była realizowana na klasycznym komputerze sekwencyjnym.

Mimo ciekawego pomysłu uzyskane wyniki przetwarzania nie prezentują się interesująco. Autorzy sami podkreślają jednak, że głównym celem było zaprezentowanie działającego systemu hybrydowego, nie zaś rozwijanie modeli systemów z kierunkowaniem uwagi (ma to być kolejny etap).

Na koniec tego podrozdziału autor pragnie wspomnieć o innych ciekawych rozwiązaniach z pogranicza sprzętu i oprogramowania. Wartymi wspomnienia wydają się tu być takie technologie jak cyfrowe procesory sygnałowe DSP (ang. *Digital Signal Processor*), układy FPGA (ang. *Field Programmable Gate Array*) czy też dedykowane układy do systemów wizyjnych takie jak procesor do estymacji ruchu STi3220.

Zakres przedstawionej powyżej tematyki wykracza jednak poza ramy tego opracowania. Osoby zainteresowane autor odsyła do literatury [24].

3.6 Zastosowania praktyczne

W niniejszej sekcji zostało zaprezentowanych kilka przykładowych zastosowań systemów z kierunkowaniem uwagi. Ze względu na ogólny charakter tejże dziedziny zastosowań jest bardzo wiele. W zasadzie FOA może być wykorzystane jako baza (preprocessing) dla niemal każdego systemu wizyjnego. W przypadku systemów wizyjnych czasu rzeczywistego wydaje się to być niemal koniecznością, szczególnie jeśli istnieje możliwość wykorzystania możliwości sprzętowej akceleracji obliczeń za pomocą dedykowanych (równoległych) jednostek obliczeniowych[7].

Jednym z pierwszych, praktycznych zastosowań systemów z kierunkowaniem uwagi była obróbka zdjęć lotniczych [42][43]. Stosując systemy FOA możliwe było przetwarzanie wielu obrazów o rozdzielczościach rzędu kilku tysięcy punktów na kilak tysięcy punktów w „rozsądnym" czasie, przy użyciu ówczesnego sprzętu[8]. Przykładowo zaprezentowany w [43] system rozpoznawał budynki na podstawie kilku zdjęć tego samego obszaru.

Bardzo praktyczne i jednocześnie proste we wdrożeniu zastosowanie FOA zaprezentowano w [37]. Autorzy pokazują jak wykorzystać FOA do znajdowania obszarów, które potencjalnie warto kompresować mniej stranie niż resztę. Dzięki takiemu podejściu uzyskany obraz zajmuje znacznie mniej miejsca na dysku lecz wizualnie („na pierwszy rzut oka") jest niemal identyczny z oryginałem. Interesującym detalem technicznym jest fakt, iż format JPEG wykorzystany przez autorów [37] posiada wbudowany mechanizm pozwalający na kompresowanie różnych obszarów z różną dokładnością, co oznacza możliwość zastosowania w/w metody bez konieczności zmiany oprogramowania po stronie klienta. W szczególności widać zastosowanie na często uczęszczanych serwisach WWW – tutaj nawet nieznaczna minimalizacja ilości danych jaką trzeba przesłać do klienta robi dużą różnicę[9].

Bazując na systemie kierunkowania uwagi stworzono także system rozpoznający na co wskazuje jego użytkownik, czyniąc niejako wirtualny interface do świata rzeczywistego [20]. System składa się z kamery oraz mini-wyświetlacza

[7] System te zostały omówione w sekc. 3.5.

[8] Należy pamiętać, iż połowa lat 90 XX wieku to epoka komputerów z procesorami klasy Pentium oraz pierwszych procesorów Pentium Pro.

[9] W praktyce wiele serwisów WWW w wersji dostępnej przez internet nie posiada formatowania (w tym również znaków końca wiersza) ani komentarzy. Przykładem takiego serwisu jest `www.google.pl`.

umieszczonych na głowie swojego użytkownika, który pokazując palcem obiekty widoczne przed sobą kieruje na nie „uwagę" systemu.

Prezentowane podejście wydaje się być ciekawym z punktu widzenia interakcji człowiek-komputer. Być może jest to preludium nowego rodzaju interface'ów, sprawiających wrażenie jakby były integralną częścią realnego świata.

W przypadku konieczności śledzenia twarzy mamy przeważnie doczynienia z systemem czasu rzeczywistego. To zaś nasuwa myśl o wykorzystaniu sytemu kierunkowania uwagi. System taki został zaprezentowany w [25]. Choć sam koncept jego działania jest ogólny system w prezentowanym zadaniu jest spójną fuzją kilku podejść. Łączy on w sobie kierunkowanie uwagi sterowane danymi i celami oraz FOA z systemem rozpoznawania.

Dzięki zastosowanemu podejściu hierarchicznemu system wykazuje dużą odporność na szum, zaś dzięki uwzględnieniu kierunkowania uwagi cechuje się dużą szybkością działania (pracuje w czasie rzeczywistym).

Dobrym przykładem jak FOA może rozwinąć istniejące już systemy jest zastosowanie takowego do aktywnego systemu monitorowania, autonomicznie śledzącego „interesujące" elementy w swoim otoczeniu [23].

Ponieważ system składa się niemal wyłącznie z elementów kierunkujących uwagę sterowaną danymi jest szybki i nadaje się do użytku praktycznego. Jeśliby go zaimplementować na dedykowanej platformie sprzętowej i zintegrować z kamerą przemysłową mógłby stanowić ciekawą alternatywę dla klasycznego monitoringu wymagającego „czujnego oka" strażnika.

Autor jednej z bardziej znanych pracy w dziedzinie FOA [27], Laurent Itti opublikował niedawno (2007 r.) dokument opisujący system rozpoznający miejsca w jakich się znajduje na podstawie zdjęć [14]. Korzysta on z podsystemu kierunkującego uwagę i na podstawie jego „trafień" tworzy niskowymiarowy (w porównaniu z obrazem wejściowym) wektor cech pozwalających na identyfikację miejsca w jakim się znajduje. Dodatkowo system współdzieli część obliczeń z FOA co zmniejsza złożoność obliczeniową całości.

Dzięki takiemu podejściu system może pracować w czasie rzeczywistym (autor mówi o przetwarzaniu do kilkunastu klastek na sekundę), co z kolei otwiera możliwości dla zastosowań w robotyce mobilnej.

Innymi przykładami zastosowań FOA w robotyce jest system VOCUS opisany dogłębnie w [45]. Choć system ten nie korzysta ze wsparcia dedykowanego sprzętu wykazuje bardzo dobre własności zarówno pod względem czasu obliczeń jak i jego właściwości.

Dobrze dostosowanym do potrzeb robotyki mobilnej jest także system zaproponowany w [34]. Jedną z poważniejszych jego zalet jest implementacja na dedykowanym sprzęcie (wysoce równoległy komputer taktowany zegarem $4[MHz]$) dającym szokująco dobre wyniki przy jednoczesnym niewielkim poboże mocy. Przyjęte w [34] podejście prawdopodobnie stanie się szybko standardem w robotyce mobilnej oraz z czasem w elektronice użytkowej.

Rozdział 4

Problemy otwarte

Niniejszy rozdział prezentuje problemy jakie pozostały nadal nie rozwiązane w zagadnieniach kierunkowania uwagi na obrazach rastrowych. Dodatkowo zawartych jest kilka pomysłów na inne rozwiązanie pewnych zagadnień, jakie zostały napotkane podczas badań literaturowych.

4.1 „Przyzwyczajanie się" uwagi

Aby kierunkowanie uwagi nie wracało ciągle w najbardziej wyróżniające się miejsce stosuje się koncept zabraniania powrotu: IOR (ang. *Inhibition Of Return*). W typowej implementacji modelu mapy występności z siecią WTA sprowadzany jest on do dezaktywacji danego obszaru na pewien okres czasu, tak aby inne fragmenty sieci mogły „wygrać" stając się nowymi punktami kierunkowania uwagi [10][27][45][34].

Aby jednak w pełni realizować ten paradygmat, w ogólnym przypadku nie wystarczy zabraniać powrotu do określonej lokacji. Takie założenie jest uproszczeniem działającym tylko w przypadku statycznych obrazów, lub sekwencji na których element przyciągający uwagę znajduje się zawsze w tym samym miejscu. Takie uproszczone podejście zawiedzie jednak kiedy proto-obiekt przyciągający uwagę będzie w ruchu. Przykładem takiego obrazu będzie plamka poruszająca się w kółko po stałej trasie na obrazie wejściowym. Tutaj podejście klasyczne nie tylko nie zadziała (obszar „zabraniany" nie będzie już zawierał proto-obiektu, który się przesunął) ale spowoduje wyłączenie na pewien czas obszaru z dalszej analizy, mimo iż potencjalnie mogłoby się tam pojawić coś co powinno przyciągnąć uwagę (np: błysk światła lub inny proto-obiekt w ruchu).

IOR musi więc mieć szerszy charakter – musi być bardziej ogólne.

4.2 Szukanie wielu celów jednocześnie

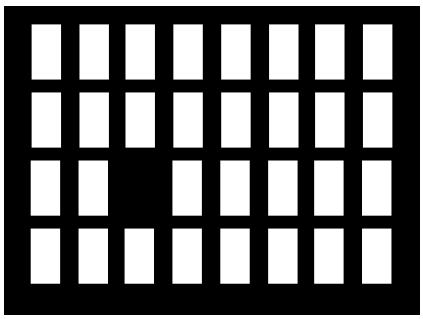
Interesujący naukowo model kierunkowania ze sterowaniem celmami przedstawiono w [45]. Zakłada on ustalanie zestawu celów jakie będą poszukiwane oraz wyuczenie systemu zestawu wag dla map widoczności, które najlepiej odzwierciedlą zadanie poszukiwania celu o zadanych właściwościach wizualnych. Potem, chcąc nakazać systemowi poszukiwanie wyuczonych proto-obiektów stosuje się odpowiedni zestaw wag do odpowiadających map.

Podejście takie dobrze zdaje egzamin kiedy mamy do czynienia z zadaniem ukierunkowania na pojedynczy cel (a w zasadzie proto-obiekt). Niestety model jest zupełnie nie skalowany w momencie kiedy zachodzi potrzeba wyszukiwania kilku celów jednocześnie. Jedynym rozwiązaniem jakie można zastosować jest osobne wyliczanie map występności na podstawie tych samych map widoczności lecz przemnożonych przez inne zestawy wag, odpowiadające innym celom. tak powstały zestaw map występności należałoby ponownie połączyć za pomocą operatora normalizacji $N(\cdot)$, tak jak to pokazano w podsekc. 3.2.3.

Z pewnością ciekawym rozwiązaniem byłoby uogólnienie modelu tak aby uwzględniał on cechy wszystkich poszukiwanych obiektów jednocześnie, bez konieczności rozbijania tego na zadania osobnego wyliczania poszczególnych celów.

4.3 Wyróżnianie poprzez brak

Jak zaobserwowano w [45], obecne systemy kierunkujące uwagę dość dobrze radzą sobie z zadaniami wyznaczania elementów wyróżniających się na tle innych pod względem pewnych cech. Nie ma jednak systemu, który pozawalałby na wyróżnienie odpowiedniego obszaru poprzez zaznaczenie braku pewnego elementu. Przykład takiej sytuacji zilustrowano na rys. 4.1. Choć każdy czło-



Rysunek 4.1: Przykład obrazu zawierającego wyróżnienie poprzez brak.

wiek od razu zauważy w tym miejscu brakujący prostokąt, systemy wizyjne będą pomijać to miejsce.

Być może uzupełnienie tego obrazu o brakujący element byłoby początkiem dla komputerowej implementacji odpowiednika domknięcia kognitywnego u ludzi.

4.4 Korzystanie z wyników poprzednich obliczeń

Choć w przypadku tworzenia wielu systemów pomija się milczeniem ich wydajność w przypadku systemów wizyjnych nie jest to poprawnym podejściem. Systemy te mają za zadanie przyspieszyć obliczenia dla algorytmów bardziej czasochłonnych (np: klasyfikacji) podając im gotowe miejsca, w których może się znajdować potencjalnie poszukiwany obiekt. Jeśli więc algorytm FOA ma spełniać swoje zadanie musi działać szybko. Jeśli będzie on zbyt wolny, może się okazać, iż proste stosowanie klasyfikatora na kolejnych wycinkach obrazu da lepsze rezultaty czasowe. Mając więc na uwadze powyższe założenia, należy zapewnić wydajność już na poziomie modelu, poprzez zaprojektowanie umożliwiające implementację na architekturze SIMD lub też będącej dostatecznie wydajnej obliczeniowo na klasycznych komputerach (zależnie od potrzeb).

Jeden z proponowanych w [45] pomysłów na skrócenie czasu obliczeń to współdzielenie obrazów wyznaczonych na etapie FOA z kolejnymi elementami systemu, tak aby nie powielać pracy. Zadanie to może w szczególności wymagać pewnych modyfikacji zarówno po stronie FOA jak i kolejnych algorytmów tak aby faktycznie korzystały one z (częściowo) tych samych przetworzonych danych.

4.5 Sterowanie celmami przez wnioskowanie

Jak wspominano w sekc. 2.3 poszczególne fragmenty mózgu nie działają w sposób zupełnie autonomiczny. Następuje pewna wymiana danych pomiędzy różnymi obszarami. W szczególności informacja o rozpoznaniu pewnego obiektu na widzianym obszarze może wpłynąć na kierunkowanie uwagi.

Obserwacja ta jest bardzo ważna i wydaje się być rozsądnym aby miała swoje odzwierciedlenie w sztucznych systemach wizyjnych. Odpowiednikiem obliczeniowym mógłby być system integrujący w sobie podsystem FOA przekazujący wyniki do podsystemu rozpoznawania, ten zaś do podsystemu wnioskowania i na koniec z niego spowrotem do FOA aby przedefiniować cele zgodnie z nowo zdobytą wiedzą. Przepływ danych w proponowanym systemie pokazuje rys. 4.2.

Rysunek 4.2: Propozycja przykładowego systemu sterowania FOA poprzez wnioskowanie o świecie na podstawie nowo zdobytej wiedzy (podsystem rozpoznawania) oraz wiedzy o świecie (podsystem wnioskujący).

4.6 FOA na różnym stopniu szczegółowości

Ludzie mają tendencję do zauważania najpierw dużych elementów, potem zaś skupiania się na szczegółach. Obecne systemy FOA nie posiadają tego rozróżnienia, choć samo wydobywanie poszczególnych elementów odbywa się osobno dla różnych skali (za pomocą piramid obrazów – patrz podsekc. 3.2.2). Być może więc dałoby się wykorzystać tę cechę jednocześnie do skrócenia czasu obliczeń (tylko obrazy w mniejszych skalach byłyby konieczne do wyliczenia od razu) i dodania etapowości kierunkowania – kiedy skonczy się przetwarzanie celów wyraźnych (dużych i wyróżniających się) system przechodziłby do dalszych etapów przetwarzając obraz z większą ilością detali (większa skala).

4.7 Wektor uwagi (uwaga przestrzenna)

Systemy kierunkujące uwagę z jakimi autor miał się okazję zapoznać można zakwalifikować z grubsza do 2 kategorii:

1. czysto wizyjne – pobierając informacje z wyłącznie jednego źródła jakim jest kamera [25][27][32].

2. mieszane – informacje o otoczeniu są pobierane za pomocą większej ilości czujników (np: kilku kamer [34] albo kamery i skamera laserowego [45]).

Pierwsza klasa problemów jest dość oczywista w sposobie analizy – istnieją dobre modle operujące na pojedynczych obrazach jako danych wejściowych [10][25].

Znacznie ciekawsza jest druga ze wspominanych klas. Większość obecnie stosowanych podejść, które zaliczają się do tej klasy korzysta z pewnego udogodnienia – dobór czujników jest tak przeprowadzony aby dane z nich wszystkich

dało się sprowadzić do pojedynczego obrazu rastrowego [45][34]. Nawet przypadek danych ze skanera laserowego omawiany w [45] został sprowadzony do obrazu rastrowego. Czujnik ten skanuje jednak fragment szerszy niż rejestruje kamera i powoduje pewne zniekształcenia geometryczne przy mapowaniu na raster $2D$ (rys. 4.3). Mimo to taki zniekształcony obraz został „nałożony" na

Rysunek 4.3: Zniekształcenia po bezpośrednim nałożeniu danych z laserowego skanera odległości na raster $2D$ – porównanie z kamerą wizyjną (zaczerpnięte z [45]).

obraz z kamery tak jakby były prezentowały one tę samą przestrzeń.

Problem ten należałoby rozwiązać jednak inaczej – potrzebna tu jest wspólna przestrzeń do reprezentowania wszystkich danych z każdego z używanych czujników.

Aby rozwiązać ten problem w sposób jak najbardziej niezależny od konkretnych typów stosowanych czujników, rozwiązanie powinno wychodzić z sedna zagadnienia – jaką przestrzeń należy przedstawić dla systemu FOA aby mógł on wskazać miejsce kierunkowania uwagi? Najrozsądniejszą odpowiedzią wydaje się tu być powierzchnia sfery. Dzięki temu, zakładając, iż wszystkie czujniki są wewnątrz tego wycinka przestrzeni, istnieje możliwość bardzo prostego nakładania kolejnych punktów kierunkowania uwagi pochodzących z różnych czujników. Proste staje się więc też odczytywanie punktu kierunkowania uwagi. Mając środek oraz znając punkt na powierzchni możemy łatwo wyznaczyć wektor kierunkowania uwagi dla systemu jako całości. Jeśli będzie taka potrzeba można także prosto sprawdzić z jakiego czujnika przyszła dana informacja, poprzez sprawdzenie, który z nich „dodał" ów punkt na powierzchni[1].

[1]W szczególności może się okazać, iż jest to grupa czujników, których przestrzeń pracy nachodzi na siebie.

Oczywiście takie rozwiązanie nie wymuszałoby na czujnikach wchodzących w jego skład niczego więcej niż możliwość podania kierunku (względem samego czujnika), z którego pochodzi sygnał. W szczególności oznacza to możliwość prostego łączenia ze sobą informacji pochodzących z takich czujników jak: wizja, zestaw mikrofonów czy czujników nacisku.

Dodatkowo prezentowany model jest elastyczny pod względem obliczeń. Umożliwia on dokonywanie wszelkich analiz na dowolnych przestrzeniach i nanoszenie wyników (wyróżniających się elementów) na powierzchnie sfery. Z drugiej strony można nanieść same sygnały na sferę i na niej wykonać wspólne obliczenia dla wszystkich czujników jednocześnie[2].

4.8 Asynchroniczne FOA i rozpoznawanie obiektów

Obecnie w istniejących systemach FOA współpracujących z podsystemami rozpoznającymi obiekty stosuje się podejście sekwencyjne – FOA wykrywa proto-obiekty, po czym są one klasyfikowane i cykl się powtarza. Wydaje się być zasadnym rozluźnienie tego wymogu. Nic nie stoi na przeszkodzie, aby FOA wyznaczało miejsca warte „rozpoznania", zaś algorytmy wyższego poziomu odpowiadałyby za ich rozpoznawanie, które może w szczególności zająć więcej niż 1 przebieg FOA. Takie uniezależnienie wprowadziłoby większą elastyczność w systemie pozwalając na niskopoziomowe przetwarzanie znacznie większej ilości klatek w tej samej jednostce czasu (np: wykrywanie sytuacji niebezpiecznych dla robota mobilnego), z drugiej zaś strony dałoby czas wyrafinowanym klasyfikatorom na wiarygodne rozpoznawanie elementów otoczenia.

4.9 Współpraca FOA z pamięcią

Ciekawym rozszerzeniem systemu wizyjnego byłoby wprowadzenie pamięci. System mógłby zapamiętywać co gdzie się znajduje w przestrzeni aby później móc wykrywać zmiany (np: brak jakiegoś elementu otoczenia, który był wcześniej). Z jednej strony jest to wyjście poza granice FOA (zewnętrzna pamięć otoczenia) z drugiej jednak taka możliwość byłaby poszerzeniem funkcjonalności oraz pewnym krokiem integrującym FOA z resztą systemu (nie tylko wizyjnego).

Byłoby to w pewnym sensie rozszerzenie nawet względem systemu wizyjnego u ludzi. Zaobserwowano bowiem, iż człowiek nie jest w stanie łatwo zauważyć różnic pomiędzy scenami, kiedy różnią się one tylko pewnymi aspektami a po-

[2]W tym przypadku będzie jednak konieczne zwiększenie wymogów odnośnie czujników – dane muszą się dać sprowadzić do wspólnej reprezentacji (np: koloru albo odległości).

między kolejnymi „prezentacjami” wstawiona jest luka czasowa [16]. System taki mógłby więc wyprzedzić pod tym względem ludzką percepcję.

4.10 „Pop-out” w wyszukiwaniu złożonym

Bardzo dobrze opisanym i jednym z podstawowych efektów kierunkowania uwagi u ludzi, jaki jest reprodukowany w sztucznych systemach, jest efekt „wyłaniania się” (ang. *pop-out*). Efekt ten działa jednak wyłącznie wtedy, kiedy obiekt różni się od pozostałych tylko pod względem dokładnie jednego atrybutu podstawowego. Wyszukiwanie złożone zawodzi zarówno u ludzi jak i w sztucznych systemach FOA w przypadku kiedy obiekt wyróżnia się większą ilością cech. Mówimy wtedy o zjawisku „wyszukiwania złożonego”.

Możliwe, iż dzięki stosunkowo prostym zabiegom udałoby się stworzyć system mający zdolność do natychmiastowego znajdywania takich elementów. Byłaby to zdecydowana poprawa względem naturalnych systemów wizyjnych.

Bibliografia

[1] http://www.usd.edu/psyc301/Rensink.htm, 2007.

[2] http://ilab.usc.edu/imgdbs/, 2007.

[3] http://www.robocup.org/, 2007.

[4] A. M. Bonnel, J. F. Stein, P. Bertucci, *Does attention modulate the perception of luminance changes?*, Quarterly journal of experimental psychology, 44A (1992).

[5] A. Reeves, G. Sperling, *Attention gating in short-term memory*, Psychological review, 93 (1986).

[6] Antonio Raffone, *Synthetic computational models of selective attention*, Neural Networks, 19 (2006).

[7] Ashutosh Sexena, Sung H. Chung, Andrew Y. Ng, *Learning depth from sinble monocular images*, (2005).

[8] B. J. Scholl, Pylyshyn, Feldman, *What is visual object? evidence from target merging in multiple object tracking.*, Cognition, 80 (2001).

[9] Bartosz Szurgot, *Zastosowanie wybranych metod sztucznej inteligencji do sterowania robotem mobilnym*, Politechnika Wprocławska, 2007.

[10] C. Koch, S. Ullman, *Shifts in selective visual attention: torwards the underlying neural circuitry*, Human neurobiology, 4 (1985).

[11] C. S. Green, D. Bavelier, *Action video game modifies visual attention*, Nature, 423 (2003).

[12] C. W. Eriksen, Y.-Y. Yeh, *Allocation of attention in the visual field*, Journal of experimental psychology: human perception and performance, 11 (1985).

[13] Chin-Hung Teng, Shang-Hong Lai, Yung-Sheng Chen, Wen-Hsing Hsu, *Accurate optical flow computation under non-uniform brightness variations*, Computer vision and image understanding, 97 (2005).

[14] Christian Siagian, Laurent Itti, *Rapid biologically-inspired scene classification with visual attention*, IEEE Transactions on pattern analisys and machine inteligence, (2007).

[15] D. LaBerge, V. Brown, *Theory of attentional operations in shape identification*, Psychological review, 96 (1989).

[16] Daniel J. Simons, Ronald A. Rensink, *Change blindness and the dynamic nature of vision*, Trends in cognitive sciences, 9 (2005).

[17] Dirk Walther, Christof Koch, *Modeling attention to salient proto-objects*, Neural networks, 19 (2006).

[18] Dominic I. Standage, Thomas P. Trappenberg, Raymond M. Klein, *Modelling divided visual attention with winner-take-all network*, Neural Netowrks, 18 (2005).

[19] Frederic Shick, Brian Scassellati, *A behavioral analisis of computational models of visual attention*, International Journal of Computer Vision, 73 (2006).

[20] Gunter Heiderman, Robert Rae, Holger Bekel, Ingo Bax, Helge Ritter, *Integrating contex-free and contex-dependent attentional mechanisms for gestural object reference*, Machine Vision and Applications, (2004).

[21] Hang Shi, Yu Yang, *A computational model of visual attention based on saliency maps*, Applied Mathematics and Computation, 188 (2007).

[22] J. G. Taylor, M. Rogers, *A control model of the movement of attention*, Neural Networks, 15 (2001).

[23] James W. Davis, Alexander M. Morison, David D. Woods, *An adaptive focus-of-attention model for video surveillance and monitoring*, Machine Vision and Applications, (2007).

[24] Kazimierz Wiatr, *Akceleracja obliczeń w systemach wizyjnych*, Wydawnictwa naukowo-techniczne, 2003.

[25] Kentaro Toyama, Gregory D. Hager, *Incremental focus of attention for robust vision based tracking*, International Journal of Computer Vision, 35 (1999).

[26] L. Carota, G. Indiveri, V. Dante, *A software-hardware selective attention system*, Neurocomputing, 58-60 (2004).

[27] Laurent Itti, *Models of bottom-up and top-down visual attention*, PhD thesis, 2000.

[28] ——, *Visual attention*. `http://ilab.usc.edu/publications/doc/Itti02hbtnn2e.pdf`, 2002.

[29] M. Bertozzi, A. Broggi, C. Caraffi, M. Del Rose, M. Felisa, G. Vezzoni, *Pedestrian detection by means of far-infrared stereo vision*, Computer vision and image understanding, 106 (2007).

[30] M. I. Posner, *Orienting of attention*, Quarterly journal of experimental psychology, 35 (1980).

[31] Maciej Huk, *Modelowanie zjawiska kierunkowania uwagi uwagi przy pomocy sztucznych sieci neuronowych w zadaniu klasyfikacji*, PhD thesis, 2006.

[32] María T. López, Miguel A. Fermández, Antonio Fernándes-Caballero, José Mira, Ana E. Delgado, *Dynamic visual attention model in image sequences*, Image and Vision Computing, 25 (2006).

[33] MaryLou Cheal, *Understanding diverse effects of visual attention with vap-filters metaphor*, Consciousness and Cognition, 6 (1997).

[34] Nabil Ouerhani, *Visual attention: from bio-inspired modeling to real-time implementation*, PhD thesis, 2003.

[35] Nabil Ouerhani, Heinz Hügli, *Real-time visual attention on a massively parallel simd architecture*, Real-time Imaging, 9 (2003).

[36] Nabli Ouerhani, Heinz Hügli, *Computing visual attention from scene depth*, (2000).

[37] Nabli Ouerhani, Javier Bracamonte, Heinz Hügli, Michael Ansorge, Fausto Pellandini, *Adaptive color image compression based on visual attention*, (2000).

[38] Neill R. Taylor, Matthew Hartley, John G. Taylor, *The micro-structure of attention*, Neural networks, 19 (2006).

[39] Patric Cavanagh, George A. Alvarez, *Tracking multiple targets with multi-focal attention*, TRENDS in Cognitive Sciences, 9 (2005).

[40] Piotr Ciesielski, Jacek Sawoniewicz, Adam Szmigielski, *Elementy robotyki mobilnej*, Wydawnictwo polsko-japońskiej wyższej szkoły technik komputerowych, 2004.

[41] Radu Horaud, David Knossow, Markus Michaelis, *Camera cooperation for achieveing visual attention*, Machine Vision and Applications, (2006).

[42] Raymond Freese, André Nel, *Focus of attention in image understanding systems*, (1993).

[43] Robert T. Collins, *Multi-image focus of attention for rapid site model construction*, (1997).

[44] Shang-Hong Lai, Baba C. Vemuri, *Reliable and efficient computation of optical flow*, International journal ofcomputer vision, 29 (1998).

[45] Simone Frintrop, *VOCUS: A visual attention system for object detection and goal-directed search*, PhD thesis, 2006.

[46] Timothée Jost, Nabil Ouerhani, Roman von Wartburg, René Müri, Heinz Hügli, *Assessing the contribution of color in visual attention*, Computer Vision and Image Understanding, 100 (2005).

[47] Vicente Concepcion, Harry Wechsler, *Detection and localization of objects in time-varying imagery using attention, representation and memory pyramids*, Pattern Recognition, 9 (1996).

[48] Yaoru Sun, Robert Fisher, *Object-based visual attention for computer vision*, Artificial Intelligence, 146 (2003).

Spis rysunków