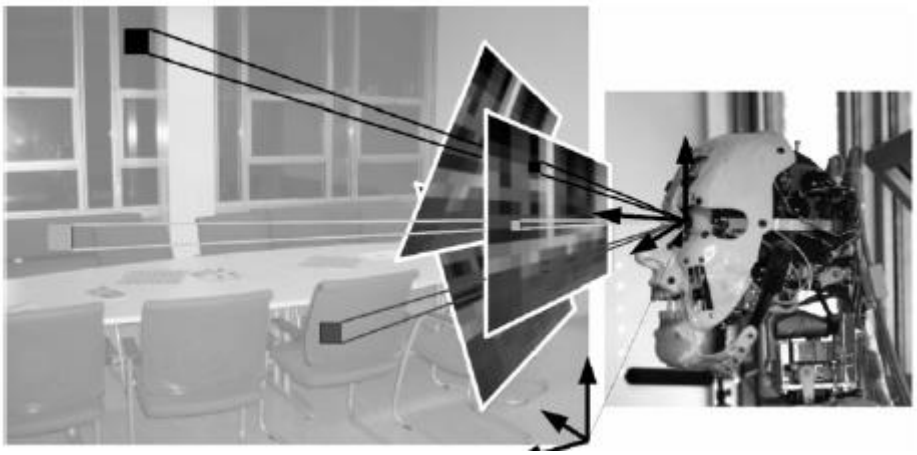
Chęć budowania sztucznych i inteligentnych systemów prowadzi do oczekiwania, że będą one działać w naszym typowym środowisku. Dlatego oczekiwania co do ich zdolności percepcyjnych są wysokie. Percepcja odnosi się do procesu uświadamiania sobie elementów środowiska poprzez doznania fizyczne, które mogą obejmować bodźce zmysłowe z oczu, uszu, nosa, języka lub skóry. W tym rozdziale skupiamy się na percepcji wzrokowej, która jest zmysłem dominującym u człowieka i wykorzystywana jest od pierwszych dni budowy sztucznych maszyn. Dwa wczesne przykłady to Shakey, mobilny robot z dalmierzem i kamerą umożliwiającą mu rozumowanie o swoich działaniach w pomieszczeniu z kilkoma obiektami, oraz FREDDY, stacjonarny robot z dwuocznym systemem widzenia, sterujący dłonią składającą się z dwóch palców. Celem widzenia komputerowego jest zrozumienie sceny lub elementów obrazów rzeczywistego świata. Ważnymi środkami do osiągnięcia tego celu są techniki przetwarzania obrazu i rozpoznawania wzorców. Analizę obrazów komplikuje fakt, że ten sam obiekt może prezentować się w aparacie na wiele różnych sposobów, w zależności od oświetlenia padającego na obiekt, kąta, pod jakim jest on oglądany, rzucanych cieni, konkretnego użytego aparatu, czy części obiektu są przesłonięte i tak dalej. Niemniej jednak dzisiejsze widzenie komputerowe jest wystarczająco zaawansowane, aby wykrywać określone obiekty i kategorie obiektów w różnych warunkach, umożliwiać pojazdowi autonomicznemu poruszanie się z umiarkowaną prędkością po otwartych drogach, kierować robotem mobilnym przez szereg biur i obserwować i zrozumieć działalność człowieka. Celem tego rozdziału jest przedstawienie aktualnego stanu wiedzy w zakresie metod widzenia komputerowego, które okazały się skuteczne i które doprowadziły do rozwoju wspomnianych powyżej możliwości. Po krótkim omówieniu bardziej ogólnych zagadnień podsumowujemy pracę podzieloną na cztery kluczowe tematy: rozpoznawanie i kategoryzacja obiektów, śledzenie i obsługa wizualna, zrozumienie ludzkich zachowań oraz zrozumienie kontekstowej sceny. Na koniec dokonamy krytycznej oceny tego, co udało się osiągnąć w zakresie wizji komputerowej i jakie wyzwania pozostają

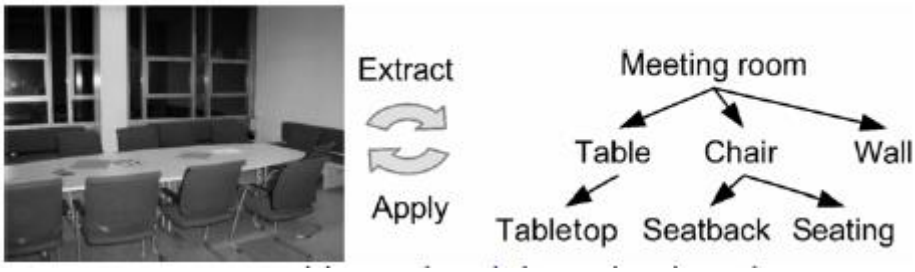
Paradygmaty i zasady widzenia komputerowego

Widzenie komputerowe to heterogeniczna dziedzina, która obejmuje szerokie spektrum metod i perspektyw naukowych. Zaczyna się od fizycznego zrozumienia, w jaki sposób powstaje obraz lub co zasadniczo można zobaczyć. Zanim światło zostanie zebrane w gęsty dwuwymiarowy układ na czujniku, zostaje załamane, odbite, rozproszone lub pochłonięte w odniesieniu do sceny. Obraz powstaje poprzez pomiar natężenia promieni świetlnych przechodzących przez każdy element układu – zwany pikselem .

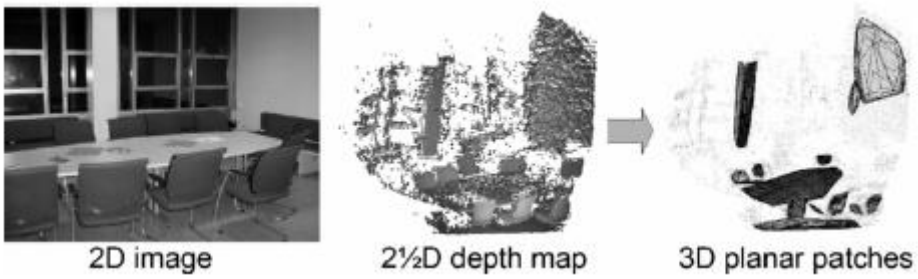


Gdyby znano oświetlenie każdego możliwego promienia światła w scenie, każdy możliwy obraz wykonany przez kamerę mógłby zostać wstępnie obliczony przed jego pomiarem. To odwzorowanie pomiędzy punktem widzenia a jego oświetleniem jest formalnie opisane przez tak zwaną funkcję

plenoptyczną. Grafika komputerowa ma na celu przybliżenie tej funkcji poprzez renderowanie znanej sceny przy danych źródłach światła. Z pierwszej perspektywy widzenie komputerowe ma na celu obliczenie odwrotnej funkcji grafiki komputerowej, to znaczy rekonstrukcję punktu widzenia i leżącej u jego podstaw sceny na podstawie danego obrazu, pary obrazów lub sekwencji obrazów. Wizja komputerowa jest tu rozumiana jako problem pomiarowy, który jest szeroko rozpatrywany za pomocą fotogrametrii, kalibracji fotometrycznej, a także technik rekonstrukcji i rejestracji. Drugie podejście do widzenia komputerowego polega na naśladowaniu widzenia biologicznego w celu głębszego zrozumienia procesów, reprezentacji i architektur. Tutaj staje się coraz bardziej oczywiste, że podstawowe pytania i otwarte problemy widzenia komputerowego znajdują się w czołówce badań nad poznaniem. Nie można ich rozwiązać w oderwaniu od rzeczywistości, lecz dotyczą one fundamentalnej podstawy samego poznania. Trzecia perspektywa rozumie wizję komputerową jako dyscyplinę inżynierską, której celem jest rozwiązywanie praktycznych zadań wizyjnych. Z jednej strony perspektywa ta domaga się wydajnych rozwiązań algorytmicznych, z drugiej zaś stawia dalsze pytanie, jak budować komputerowe systemy wizyjne. Stan wiedzy w tej dziedzinie jest zdominowany głównie przez heurystykę i wiedzę płynącą z doświadczenia. Systematyczne podejścia metodologiczne są rzadkie, w większości specyficzne dla aplikacji i obecnie brakuje im głębokiego zrozumienia problemu wzroku jako takiego. Tym samym nie da się rozdzielić wszystkich trzech perspektyw i głęboko na siebie oddziaływać, co – w połączeniu z ogromnym postępem technicznym – sprawiło, że wizja komputerowa stała się w ciągu ostatnich pięćdziesięciu lat dziedziną niezwykle dynamiczną. Aby rozwiązać określone zadania związane z wizją komputerową, należy podjąć różne decyzje projektowe. Niektóre z nich wskazano poniżej. Jaka wiedza jest potrzebna? Aby zrozumieć treść obrazu, odpowiednie jego części muszą zostać powiązane z koncepcjami znaczącymi semantycznie. W przypadku sceny sali konferencyjnej

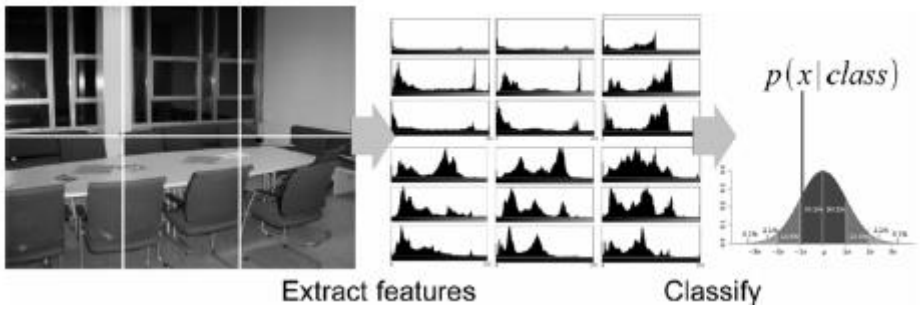


baza wiedzy może zawierać informację, że składa się ona z dużego stołu i kilku ustawionych wokół niego krzeseł, że stół ma blat i tak dalej. Baza wiedzy rozkłada złożoną scenę sali konferencyjnej na prostsze elementy, takie jak blat, które odpowiadają płaskiej powierzchni lub jednorodnemu obszarowi, który można bezpośrednio wyodrębnić z obrazu. Dlatego algorytm może rozpocząć się od przeszukania obrazu pod kątem jednorodnych obszarów, co jest koncepcją niskiego poziomu w odniesieniu do sygnału. Następnie są one sukcesywnie łączone (w oparciu o bazę wiedzy) w celu utworzenia koncepcji wyższego poziomu. Podejście to jest zwykle określane jako „oddolne". Inny algorytm może zacząć od koncepcji tabeli i szukać konkretnie konfiguracji części (przewidywanej przez bazę wiedzy), która spełnia wymagania tej koncepcji. Te części z kolei mogą aktywować detektor stołowy, który jest nakładany na obraz. Podejście to jest zwykle określane jako „odgórne". Obydwa podejścia do bazy wiedzy pomogły w prowadzeniu znacznych badań nad wizją komputerową w latach 70. i 80. XX wieku. Jak przedstawić geometrię sceny? Geometria sceny jest ważna jako reprezentacja pośrednia w procesie interpretacji obrazu. Można to rozwiązać w trybie 2D lub 3D. Na rysunku scena jest przedstawiona jako zwykły obraz 2D (po lewej) i obraz głębi (pośrodku).

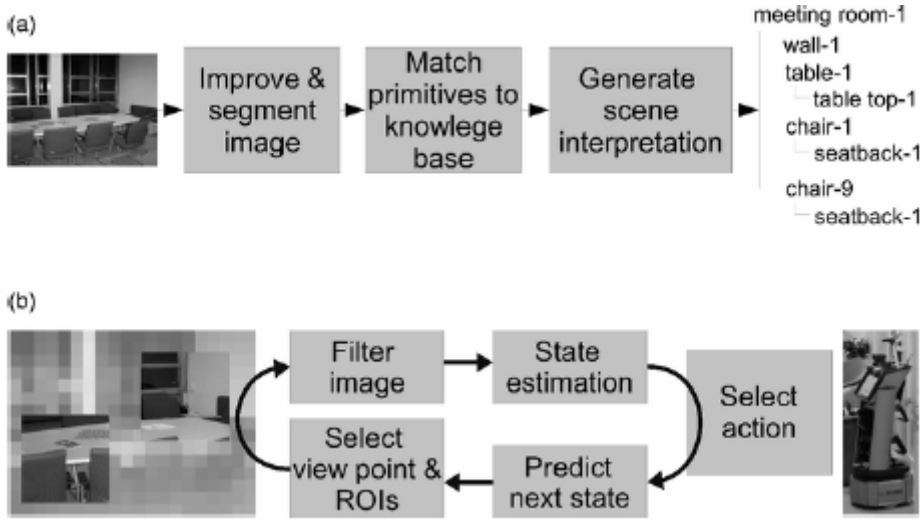


Tę ostatnią można obliczyć na podstawie par obrazów stereo lub bezpośrednio zmierzyć, na przykład za pomocą czujników czasu przelotu, które mierzą odległość do każdego piksela poprzez modulację i odbieranie wiązki światła podczerwonego. Ponieważ reprezentacja we współrzędnych pikseli nadal zależy od widoku, nazywa się ją również 2½D. W następnym kroku do sceny dopasowywane są geometryczne prymitywy 3D, przy czym każde dopasowanie definiuje transformację geometryczną. Ponieważ teraz znane jest względne położenie 3D i orientacja 3D pomiędzy tymi prymitywami, osiągnięta została reprezentacja niezależna od widoku i skupiona na obiekcie. Podejście tego typu zostało pierwotnie zaproponowane przez Davida Marra, który zajął się także znanymi wówczas koncepcjami ludzkiego wzroku. Jednak w wielu przypadkach ekstrakcja geometrii 3D jest zbyt delikatna. Kształty rzeczywistych obiektów 3D są często złożone i niesztywne, a procedury dopasowywania często kończą się lokalnymi minimami oraz błędnym położeniem i orientacją obiektu („pozą"). W rezultacie z obrazów 2D można również uzyskać bardziej stabilne reprezentacje geometryczne. W tym przypadku obrazy analizowane są pod kątem nieciągłości przestrzennych w poziomie szarości lub powierzchni barwnej. Reprezentacje skupiają się albo na jednorodnych obszarach obrazu (obszarach), albo na krawędziach (liniach granicznych) .

Obydwa stanowią podstawę do dalszych procesów interpretacyjnych. Wyodrębnianie takich prymitywów geometrycznych stanowi problem cyfrowego przetwarzania obrazu. Jakie są odpowiednie cechy? Aby dopasować reprezentację geometryczną lub obrazową do koncepcji semantycznej, takiej jak „stół", „krzesło" lub „sala konferencyjna", należy określić funkcję decyzyjną, która decyduje o przynależności do klasy lub przeciw niej. Jest to problem klasyfikacyjny, który jest intensywnie poruszany w obszarze rozpoznawania wzorców . Wzór jest reprezentowany przez wektor cech określający punkt w przestrzeni wielowymiarowej. Biorąc pod uwagę, że znane są klasy niektórych punktów w tej przestrzeni (np. zbiór obrazów treningowych z odręcznymi adnotacjami), można nauczyć się funkcji decyzyjnej, która dzieli przestrzeń na te klasy. Na rysunku przedstawiono prosty przykład.



Obraz jest dzielony na sześć części i dla każdego podobrazu obliczany jest histogram kolorów. Połączone histogramy stanowią wektor cech, który można wykorzystać na przykład do klasyfikacji określonych sal konferencyjnych. Pytanie, jakie są dobre cechy, jest tematem od dawna dyskusja. Na przestrzeni lat pojawiło się kilka wynalazków, które wywarły ogromny wpływ na tę dziedzinę. W latach 90. Swain i Ballard zaproponowali wykorzystanie statystyk cech lokalnych (takich jak histogramy kolorów), Turk i Pentland zastosowali technikę opartą na wektorach własnych do obrazowania zbiorów ludzkich twarzy (zwanych wówczas twarzami własnymi). Później, w 2000 roku, Viola i Jones zrewolucjonizowały wykrywanie twarzy, wymyślając proces automatycznego wyboru cech w oparciu o ogromną liczbę bardzo prostych cech związanych z falkami Haara (funkcje oparte na binarnym włączaniu/wyłączaniu sąsiednich części obrazu). Kolejnym przełomem była transformacja cech niezmiennych skali (SIFT) autorstwa Davida Lowe'a, która wyniosła rozpoznawanie obiektów na nowy poziom. Tutaj idee lokalnej statystyki gradientów łączą się z wyjątkowo stabilną detekcją stałych punktów na obiekcie – tak zwanych punktów procentowych. Jak kontrolować proces akwizycji? Widzenie biologiczne nie jest pasywnym procesem interpretacji i nie powinno tak być w przypadku autonomicznych sztucznych systemów. Poruszanie się agenta w świecie rzeczywistym w zasadzie determinuje problem percepcji, jaki musi on rozwiązać. Wizja rozumiana jest jako proces aktywny obejmujący kontrolę czujnika i ściśle powiązany z pomyślną realizacją decyzji lub działania . Ma to pewne konsekwencje dla projektowania komputerowych systemów wizyjnych, co zauważono już na początku lat 90. XX wieku . Po pierwsze, zamiast modelować izolowany proces interpretacji obrazu, system musi być zawsze uruchomiony i musi kontrolować swoje zachowanie za pomocą strumienia obrazu. Po drugie, ogólnym celem przetwarzania wizualnego nie jest zrozumienie obrazu. Zamiast tego system wizyjny musi działać jak filtr wydobywający informacje istotne dla jego zadania. Po trzecie, system musi reagować z określonym opóźnieniem, aby był przydatny w bieżącym zadaniu, takim jak nawigacja i omijanie przeszkód w robocie. Po czwarte, zamiast przetwarzać cały obraz, system musi skoncentrować się na obszarze zainteresowania (ROI), aby osiągnąć cele w zakresie wydajności. Różne perspektywy pokazano na rysunku.



Pierwsza ma na celu pełną interpretację obrazu, druga wydobywa istotne informacje do wyboru działań i przewidywania stanu.

Rozpoznawanie i kategoryzacja obiektów

Rozpoznawanie obiektów można postrzegać jako wyzwanie polegające na określeniu „gdzie" i „co" obiektów w scenie. Zaproponowano wiele różnych technik i wszystkie mają swoje zalety i wady. Biorąc pod uwagę scenariusz zastosowania, należy starannie wybrać odpowiednią technikę rozpoznawania obiektów, która spełnia przewidywany zestaw ograniczeń. Techniki różnią się także dokładnym problemem, który rozwiązują. Wiele technik rozpoznawania to detektory obiektów, które zadają pytanie „tak/nie" dotyczące obecności klasy obiektów. Obraz jest zazwyczaj skanowany za pomocą modelu szablonowego; oznacza to, że nad obrazem przesuwane jest okno i dla każdej pozycji obliczana jest tzw. odpowiedź filtra poprzez dopasowanie szablonu do podobrazu zdefiniowanego przez okno. Każda inna parametryzacja obiektu (skala obiektu, obrót itp.) wymaga osobnego skanowania. Bardziej wyrafinowane podejścia skutecznie wykonują wielokrotne przejścia w różnych skalach i stosują filtry wyuczone na podstawie dużych zestawów oznaczonych obrazów. Dobrym przykładem jest wspomniany w poprzednim rozdziale detektor twarzy firmy Viola and Jones. W tym przypadku filtr składa się ze zbioru całek dodatnich i ujemnych po poznanych wcześniej prostokątnych obszarach obrazu. Techniki oparte na segmentacji najpierw wyodrębniają opis geometryczny obiektu poprzez grupowanie pikseli, które definiują rozszerzenie obiektu na obrazie. Jest to typowy proces oddolny, jak omówiono wcześniej. W drugim etapie techniki te obliczają niezmienny zestaw cech. Właściwość niezmienności oznacza, że cechy zachowują te same lub podobne wartości w przypadku różnych przekształceń obrazu, takich jak skalowanie, obracanie lub zmiana oświetlenia. Następnie funkcje są wykorzystywane do rozpoznawania klasy obiektów lub wyodrębniania zestawu ogólnych prymitywów, z których zbudowane są obiekty. Nowoczesne techniki przeplatają lub łączą oba etapy, aby uporać się z problemami nadmiernej segmentacji (w której części są dzielone na małe kawałki) i niedostatecznej segmentacji (w której części są grupowane razem z obszarami tła). Metody dopasowywania wykorzystują „parametryczne" modele obiektów dopasowywane do danych obrazu. Algorytm musi szukać parametrów takich jak skalowanie, obrót czy translacja, które optymalnie dopasowują model do odpowiednich cech obrazu. Przybliżone rozwiązanie można znaleźć także poprzez proces odwrotny, czyli cechy obrazu (np. narożniki, kontury czy inne charakterystyczne punkty obrazu) głosują na rozwiązania parametrów zgodne z wykrytą cechą (proces polega na wykorzystaniu schematu głosowania lub algorytm, który wyprowadza pojedynczy wynik z wielu źródeł danych). W tym przypadku przestrzeń parametrów jest zgrubnie dyskretyzowana. Technika ta jest często określana jako uogólniona transformata Hougha , a jej wariant został zastosowany w rozpoznawaniu obiektów

przez Davida Lowe'a, o którym mowa w ostatniej sekcji . Wszystkie trzy podejścia dostarczają różnych informacji o obiektach na obrazach i zakładają, że dostępne są różne rodzaje wiedzy wstępnej.

Modelowanie 2D

Większość obiektów w świecie rzeczywistym jest z natury trójwymiarowa. Niemniej jednak wiele technik rozpoznawania obiektów ze znacznym sukcesem opiera się na reprezentacjach 2D. Jest tego kilka powodów. (1) Łatwa dostępność: Informacje o obrazie 2D uzyskujemy niemal za darmo, korzystając ze standardowego sprzętu fotograficznego. (2) Szybkie obliczenia: Cechy można obliczyć bezpośrednio na podstawie danych pikseli obrazu i nie wymagają one wyszukiwania skomplikowanych prymitywów geometrycznych. (3) Proste pozyskiwanie modeli detekcji: Modele używane do automatycznego wykrywania obiektów są zazwyczaj uczone z przykładowych obrazów. (4) Odporność na szum: Cechy są obliczane bezpośrednio na wartościach pikseli. Kontrastuje to z ekstrakcją bardziej abstrakcyjnych prymitywów (regionów, konturów, prymitywów kształtów 3D), która zazwyczaj wiąże się z problemami z segmentacją i dlatego jest bardziej podatna na błędy w odniesieniu do bałaganu i szumu. (5) Ponadto wiele interesujących obiektów ma dość charakterystyczne widoki 2D – na przykład strony tytułowe, znaki drogowe, widoki motocykli lub samochodów z boku, widoki twarzy z przodu. Ceną, jaką trzeba zapłacić za ignorowanie cech 3D obiektów, są zazwyczaj modele nadmiernie lub niedostatecznie ograniczone, ponieważ istnieje wiele odmian perspektywy, z którymi nie można sobie poradzić w sposób systematyczny. Typowym przypadkiem podejść z niedostatecznymi ograniczeniami są modele zbioru funkcji. Podobnie jak modele histogramów, obliczają one statystyki funkcji dla obszaru obrazu lub całego obrazu. W ten sposób lokalizacja obiektów zostaje całkowicie utracona i nie można rozróżnić obrotu obiektu ani jego dokładnego położenia. Tak więc, na przykład, jeśli oczy, nos i usta twarzy były odwrócone do góry nogami lub całkowicie pomieszane, moduł rozpoznawania nadal błędnie wykrył twarz. Z drugiej strony modele nadmiernie powiązane wymagają wielu reprezentacji, aby poradzić sobie z różnymi konfiguracjami części lub obrotami obiektów. (Dobrym przykładem są wspomniane wcześniej metody oparte na szablonach.) Zatem jeśli na przykład twarz zostanie obrócona o 90 stopni, moduł rozpoznawania nigdy jej nie wykryje. Za dodatkową cenę musimy uporać się z trudniejszym problemem segmentacji – czyli problemem wyodrębnienia obiektu z tła. Zazwyczaj tło jest dalej, więc informacje 3D zapewniają znacznie silniejszą wskazówkę niż wartości luminancji obrazów 2D. Dominującą klasą technik rozpoznawania obiektów 2D są podejścia oparte na wyglądzie. Zamiast używać niezmiennej względem widoku reprezentacji skupionej na obiekcie, reprezentują one różne aspekty obiektu. Zwarte reprezentacje są dostarczane przez wykresy aspektowe , które łączą ze sobą różne wyglądy 2D w wydajnej strukturze danych. Po drugie, podejścia oparte na wyglądzie obniżają pośredni poziom reprezentacji geometrycznej poprzez obliczanie cech bezpośrednio z wartości pikseli. Ma to pewne konsekwencje dla rodzaju klas obiektów, które można wyróżnić, i odmian wewnątrzklasowych, które można uwzględnić. Jak dotąd omówione metody zajmują się zmianami obrotu, oświetleniem, hałasem i niewielkimi zniekształceniami kształtu obiektu. Zakładają przeważnie, że obiekty są solidne, w przybliżeniu sztywne, mają podobną teksturę lub kolor i są w niewielkim stopniu przesłonięte. Dalsze odmiany są objęte podejściami opartymi na lokalnych deskryptorach. W tym przypadku główną ideą jest wykrycie najistotniejszych punktów na obrazie, które zapewniają częściowy opis funkcji, a nie pełny model wyglądu. Podejścia te zwróciły na siebie uwagę w pierwszej dekadzie XXI wieku i osiągnęły niespotykany wcześniej poziom. Opierając się na lokalnych deskryptorach (typowymi przykładami są funkcje SIFT lub SURF, które analizują rozkład gradientów obrazu wokół punktu obrazu), metody te są w stanie poradzić sobie z okluzją i lokalnymi zmianami, które występują w warunkach rzeczywistych.

Modelowanie 3D

Obrazy 2D o kolorze lub intensywności nie kodują bezpośrednio informacji o głębi ani kształcie. W konsekwencji rozpoznawanie i lokalizacja obiektów jest problemem trudnym i generalnie źle postawionym . Aby przezwyciężyć te problemy, kształt 3D obiektów można bezpośrednio odtworzyć z obrazów głębi lub zakresu. Obrazy głębi można uzyskać różnymi metodami, począwszy od skanowania za pomocą czujnika laserowego, przez metody wykorzystujące światło strukturalne, po systemy stereofoniczne wykorzystujące dwie kamery, czyli metodę stosowaną przez ludzki wzrok. Tanim przykładem kamery ze światłem strukturalnym jest kamera Kinect obsługująca kolor i głębię obrazu. Głównym pytaniem w wizji komputerowej jest to, jak modelować lub reprezentować obiekt w taki sposób, aby można go było wykryć w danych szczegółowych. Jednym ze sposobów jest rozbicie kształtów na części składowe i zdefiniowanie ich relacji przestrzennych. W wizji komputerowej części są przydatne z dwóch powodów. Po pierwsze, wiele obiektów jest artykułowanych, a opis oparty na częściach pozwala nam oddzielić kształty części od ich relacji przestrzennych. Po drugie, nie wszystkie części obiektów są widoczne, ale często wystarczą, aby rozpoznać obiekt; na przykład filiżankę można rozpoznać po korpusie lub rączce. Kluczowym aspektem reprezentacji opartych na częściach jest ich liczba parametrów. W ostatniej dekadzie włożono wiele pracy w opisanie danych dotyczących głębokości za pomocą prymitywów rotacyjno-symetrycznych (kula, walec, stożek, torus). Uogólnione cylindry można utworzyć poprzez przeciągnięcie konturu 2D wzdłuż dowolnej krzywej przestrzennej . Ponieważ kontur może zmieniać się wzdłuż krzywej (osi), potrzebne są definicje osi i krzywej odchylenia, aby zdefiniować uogólniony cylinder, co wymaga dużej liczby parametrów. Często cytowanym wczesnym systemem wizyjnym wykorzystującym uogólnione cylindry jest system ACRONYM do wykrywania samolotów . Dopasowanie wielu parametrów jest jednak skomplikowane i ogranicza zastosowanie tej metody. Jedna z najczęściej badanych metod modelowania 3D polega na odzyskiwaniu superkwadryk – kształtów geometrycznych zdefiniowanych za pomocą wzorów obejmujących dowolne potęgi w celu wytworzenia kształtów przypominających sześciany, cylindry i stożki, z zaokrąglonymi lub ostrymi narożnikami. Stały się one popularne, ponieważ mały zestaw parametrów może opisać szeroką gamę różnych podstawowych kształtów. Solina i in. był pionierem prac nad odzyskiwaniem pojedynczych superkwadryk i wykazał, że odzyskiwanie superkwadryk z obrazów zakresowych jest wrażliwe na szum i wartości odstające, w szczególności z pojedynczych widoków stosowanych w zastosowaniach takich jak robotyka. Jaklic i współpracownicy podsumowują paradygmat odzyskiwania i selekcji służący do segmentowania sceny za pomocą prostych obiektów geometrycznych bez okluzji. Metoda ta ma na celu pełne wyszukiwanie z otwartym czasem przetwarzania nieodpowiednim dla większości zastosowań, takich jak robotyka. Ostatnio do uzyskania danych 3D coraz częściej wykorzystuje się obrazy z czujników głębi, takich jak Kinect lub z systemów stereo. Ponieważ dane nie są na ogół tak dobre, jak ze skanów laserowych, stosuje się metody statystyczne, a nie metody bezpośredniego kształtowania. Otwartymi problemami w tym obszarze są: radzenie sobie z rzadkimi danymi wynikającymi ze skanów sceny w jednym widoku, radzenie sobie z typowymi cieniami i okluzjami lasera i kamery w zagraconych scenach oraz radzenie sobie z niepewnością obrazów stereo.

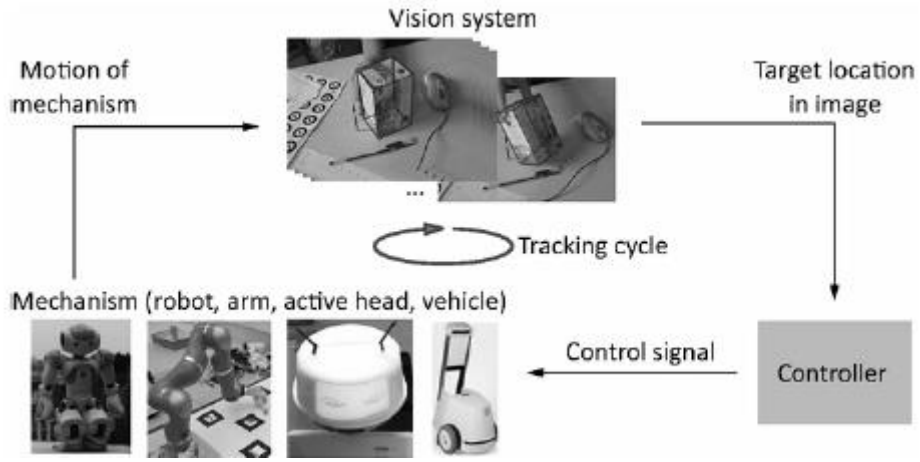
Śledzenie i obsługa wizualna

Innym typowym zadaniem człowieka jest wykrywanie i śledzenie ruchu obiektów. Podczas chwytania przedmiotu obserwuje się ruch względny. Podczas chodzenia monitorowany jest ruch otoczenia. Technika wizualnego śledzenia obiektu i określania jego lokalizacji znajduje zastosowanie szczególnie w zadaniach inwigilacyjnych i robotyki. W pierwszym przypadku szacuje się, że ścieżki samochodów lub osób przywrócą bieżące działania i odpowiednio zareagują. W robotyce celem jest śledzenie względnej pozycji robota mobilnego i jego otoczenia lub skierowanie ręki robota w stronę obiektu. Ciągła kontrola zwrotna pozycji robota nazywana jest serwowaniem wizualnym. Pierwsze sukcesy w prowadzeniu pojazdów autonomicznych i prowadzeniu pojazdów powietrznych wskazują na

zastosowanie serwowania wizualnego . Jednakże nadal istnieją dwie główne przeszkody w dalszym wykorzystaniu tej metody w rzeczywistych scenariuszach. Po pierwsze, wymagany jest wydajny cykl śledzenia. Aby zapewnić dobrą dynamikę, należy połączyć wizję i kontrolę. Szybkie ruchy są potrzebne, aby uzasadnić zastosowanie serwonapędu wizualnego w rzeczywistych zastosowaniach robotycznych. Po drugie, musi istnieć niezawodne wykrywanie obiektu docelowego. Wizja musi być solidna i niezawodna. Percepcja musi być w stanie ocenić stan obiektów i robota, umożliwić robotowi reakcję na zmiany i zapewnić bezpieczne poruszanie się w swoim otoczeniu. Zagadnieniu cyklu śledzenia poświęcono wiele uwagi w literaturze, ale skuteczne wykrywanie celów wizualnych jest równie istotne i ostatnio zaczęło poświęcać mu coraz więcej uwagi. W poniższych sekcjach podsumowano stan wiedzy w odniesieniu do tych dwóch kryteriów

Cykl śledzenia

Celem serwowania wizualnego jest uwzględnienie całego systemu i jego interfejsów. Podstawową pętlę sterowania przedstawiono na rysunku.



Zawiera trzy główne bloki: system wizyjny, kontroler i mechanizm (lub robot lub pojazd). System wizyjny określa aktualne położenie celu (obiektu zainteresowania) na obrazie. Sterownik konwertuje lokalizację na obrazie na pozycję w przestrzeni lub bezpośrednio na wartości poleceń. System powtarza to z częstotliwością cyklu. W każdym cyklu wyznaczana jest nowa lokalizacja, możliwe jest także wykorzystanie różnicy lokalizacji w celu uzyskania polecenia sterującego. Robot lub pojazd zwykle wykorzystuje oddzielny kontroler do sterowania silnikami na poziomie osi i kół. Celem jest zbudowanie systemu śledzenia w taki sposób, aby cel nie został zgubiony. Jedynym ograniczeniem śledzenia jest pole widzenia kamery. Dlatego przydatne jest zbadanie śledzenia najwyższej możliwej prędkości celu (lub przyspieszenia). Odpowiednią właściwością jest opóźnienie (lub latencja) informacji zwrotnej generowanej przez system wizyjny. Dwa główne czynniki, na które należy zwrócić uwagę, to (1) opóźnienie lub opóźnienia w jednym cyklu oraz (2) część lub okno obrazu, które jest faktycznie przetwarzane. Opóźnienia z kamery kumulują się. Obecnie kamery wytwarzają obrazy z częstotliwością 25 lub 30 Hz lub obrazy na sekundę. Dodatkowe opóźnienia wynikają z czasu potrzebnego na przesłanie danych obrazu do kontrolera. Największe opóźnienie czasowe to czas potrzebny na przetworzenie obrazu. Chociaż wydaje się intuicyjne, że opóźnienia opóźniają śledzenie, drugi czynnik, czyli przetwarzanie obrazu, często nie jest przestrzegany. Jeśli obliczony zostanie pełny obraz, może to zająć znacznie więcej czasu niż czas rejestrowania obrazu w aparacie, co może skutkować utratą obrazów. Jeśli zastosuje się małe okno, na przykład wokół miejsca, w którym widziano cel na ostatnim obrazie, możliwe jest wykorzystanie każdego obrazu. Optimum osiąga się, gdy rozmiar okna jest tak dobrany, że przetwarzanie jest tak szybkie, jak pozyskiwanie obrazów, a przetwarzanie obrazu odbywa się przy

tej samej częstotliwości 25 lub 30 Hz . Oznacza to, że optymalne jest stosowanie systemu śledzenia z opóźnieniem wynoszącym dwa cykle szybkości klatek dla kamer: jeden do przesyłania obrazu z kamery do komputera, a drugi do przetwarzania obrazu. Aby skompensować to opóźnienie, filtry (takie jak filtr Kalmana) przewidują, gdzie będzie cel. Warto zauważyć, że ludzkie oko bardzo różni się od aparatu. Kamery mają jednolity układ pikseli przy danej rozdzielczości lub odstępie między pikselami. Ludzka siatkówka wykazuje przestrzenną teselację z dołkiem o wysokiej rozdzielczości pośrodku i szerokim polem widzenia (około 180 stopni) przy logarytmicznie malejącej rozdzielczości. Efekt jest taki, że człowiek cały czas przetwarza cały obraz . Ludzie mogą reagować na ruch na obrzeżach, podczas gdy rozpoznawanie działa tylko w dołku, który jest obrócony w stronę celu i śledzi go.

Solidne wykrywanie celu

Solidność śledzenia jest głównym problemem w zapewnieniu ciągłej pracy w aplikacjach. Powiedzieć, że metoda śledzenia jest solidna, oznacza, że ulega ona płynnej degradacji, gdy dane wejściowe są zaszumione i zawierają wartości odstające. Wspólnym mianownikiem technik zwiększających niezawodność jest wykorzystanie redundancji poprzez użycie wielu kamer, wielu rozdzielczości, ograniczeń czasowych właściwych dla śledzenia, modeli i integracji kilku wskazówek lub funkcji. Minimalna forma redundancji jest nieodłączną cechą systemu stereowizyjnego wykorzystującego dwie stałe kamery i poszukującego celu na obu obrazach. Obecnie na rynku dostępne są systemy obliczające obraz głębi z dwóch obrazów stereo. Niemniej jednak problem zgodności (znalezienie tego samego punktu sceny na obu obrazach) pozostaje nadal, a udane zastosowania stereo są rzadkie. Problem zgodności widzenia stereoskopowego zostaje zredukowany dzięki zastosowaniu trzech lub więcej kamer, jak w TRICLOPS (Point-Grey Research). Systemy wspomagające do kierowania samochodami przy dużych prędkościach wykorzystują dwie lub trzy kamery o różnych polach widzenia. Pomysł łączenia informacji z różnych poziomów rozdzielczości został wykorzystany w podejściach skali-przestrzeni lub piramidy obrazu, gdzie rozmiar oryginalnego obrazu jest kilkakrotnie zmniejszany. Spójność jest agregowana na mniejszych obrazach, aby uzyskać miarę niezawodności, na przykład wykrywania krawędzi. Ostatnio cechy punktów procentowych (cechy, które mają maksymalne gradienty) wykorzystują to do wybrania najbardziej niezawodnej skali lokalnej punktu gradientu, na przykład SIFT . Jednak wykorzystanie piramid obrazowych nadal nie jest dostatecznie wykorzystywane. Redundancję serii obrazów można wykorzystać, biorąc pod uwagę czasową spójność wykrytych cech, zwaną także czasowym powiązaniem danych. Aby poradzić sobie z niepewnością związaną z lokalizacją obiektu docelowego na obrazie, powszechnie stosuje się standardowe metody teorii sterowania, takie jak filtrowanie i przewidywanie , w celu poprawy odporności. Obecnie najpowszechniejszym podejściem do radzenia sobie z tą niepewnością jest filtrowanie Kalmana lub filtrowanie cząstek, w przypadku którego kilka hipotez pomaga w dostosowaniu się do niepewności ruchu mechanizmu i pomiaru. Podejście oparte na wizji dynamicznej wykorzystywało czasową ewolucję cech geometrycznych, takich jak linie, do zbudowania modelu postrzeganego świata. Właściwości fizyczne obiektów, takie jak pewna bezwładność, są wykorzystywane do przewidywania przyszłych pozycji obiektu na kolejnych obrazach. Śledzenie służy następnie do potwierdzenia lub aktualizacji modelu ruchu. Innym podejściem jest wizja oparta na modelach. Model jest zwykle reprezentacją celu w formacie CAD (projektowanie wspomagane komputerowo), która służy do przewidywania położenia obiektu (modelu) na następnym obrazie. Roboty mobilne przechowują (lub budują) reprezentacje obiektów, takich jak ściany, filary lub pudełka, do celów nawigacji lub chwytania obiektów. U ludzi zintegrowanie wskazówek lub cech, takich jak tekstura, kolor, cieniowanie itd., zostało zidentyfikowane jako prawdopodobne źródło doskonałej zdolności radzenia sobie ze zmieniającymi się warunkami. Podsumowując, istnieje mnóstwo podejść do śledzenia. Większość z nich jest solidna lub szybka. Podczas gdy śledzenie oparte na regionach lub punktach szczególnych jest bardziej niezawodne w środowiskach teksturowanych, schematy śledzenia oparte na krawędziach zapewniają najlepsze

dane wejściowe do serwowania wizualnego w robotyce lub systemach rzeczywistości rozszerzonej, gdzie dodatkowe informacje są wizualizowane na rzeczywistych obrazach. Wraz ze stałym wzrostem mocy obliczeniowej prace nad integracją sygnałów będą posuwać się dalej. Można wiele osiągnąć, wykorzystując większą wiedzę na temat zadania i dziedziny, modeli obiektów i funkcji obiektów, a także korzystając ze wskazówek, takich jak poziomy rozdzielczości, spójność czasowa i różne cechy obrazu.

Zrozumienie ludzkich zachowań

Nadzór wizualny

Inteligentne pokoje, interfejsy człowiek-maszyna oraz aplikacje związane z bezpieczeństwem i ochroną wymagają umiejętności rozpoznawania działań ludzi. Dziedzina ta nazywana jest nadzorem wizualnym. Zazwyczaj systemy monitoringu działają w oparciu o kamery stacjonarne, co pozwala na zastosowanie techniki odejmowania tła w celu wykrycia zmian w obrazie. Odejmowanie tła wykorzystuje obrazy statyczne w celu uzyskania modelu nieruchomej sceny tła, co upraszcza zadanie wyodrębniania poruszających się obiektów na pierwszym planie (pojazdów, osób itp.) Głównym zadaniem jest radzenie sobie ze zmiennym oświetleniem, które zmienia wygląd obrazu i może ukryć zmiany spowodowane poruszającymi się obiektami na pierwszym planie. Ta forma wykrywania zmian skutkuje powstaniem obszarów obrazu, które służą jako wskazania obiektów. W następnym kroku te plamy są śledzone w sekwencji obrazów, gdzie wykorzystywane są metody łączenia danych w celu znalezienia stale poruszających się obiektów i wykrycia wygenerowanych błędnych obszarów. Preferowanymi metodami modelowania stale poruszającego się obiektu są ukryte modele Markowa i sieci Bayesa. Systemy nadzoru często działają w dwóch fazach: fazie uczenia się i fazie działania. W fazie uczenia się system jest inicjowany w danej scenie, a modele są dostosowywane lub uczone na podstawie obserwacji. Modele te zawierają dane dotyczące normalnych czynności, takich jak pasy ruchu samochodów, punkty wejściowe czy typowe ludzkie gesty. W fazie wykonawczej strumienie danych porównuje się z danymi modelu, aby uzyskać interpretacje i reakcje. Obecnie systemy potrafią wykrywać i rozpoznawać zachowania od kilku osób aż po większe grupy ludzi . W scenach ruchu drogowego przetwarzanie odbywa się głównie oddolnie, podczas gdy nowsze systemy wykorzystują wiedzę domenową w sposób odgórny . Przykładem jest wykorzystanie modeli obiektowych i modeli oczekiwanej aktywności do monitorowania aktywności na płytach postojowych . W dziedzinie robotyki relację obiekt-człowiek badano w podejściach takich jak programowanie przez demonstrację (PbD), gdzie zadaniem jest interpretacja poleceń użytkownika w celu nauczenia robota. W PbD użytkownik albo fizycznie prowadzi ramię robota przez ruch, albo system wizyjny rejestruje ruch ludzkiego ramienia i przenosi go na ramię robota. W najnowszej pracy czynności rąk i przedmiotów są interpretowane i zapisywane przy użyciu wyrażeń języka naturalnego w planie zajęć – zwięzłym opisie scenariusza określającym odpowiednie obiekty i sposób, w jaki się na nie reaguje. Wraz ze spadkiem kosztów kamer obecny kierunek prac zmierza w kierunku sieci kamer obsługujących duże obszary. Szczegółowe modele ludzi i ich typowych czynności pozwalają na lepszą interpretację gestów w mniej ograniczonych warunkach

Interakcja człowiek-maszyna

Przejście od technik obserwacji wizualnej do opartego na wizji interaktywnego interfejsu człowiek-komputer wydaje się być małym krokiem. Otwiera pełną gamę nowych zastosowań, w których komputery, monitory i urządzenia wejściowe, takie jak klawiatura i mysz, znikają z codziennego otoczenia. Na przykład prosty gest dłoni i spojrzenie mogą przenieść kolekcję zdjęć z aparatu na duży ekran telewizora w salonie, zamieniając ludzkie ciało w kontekstowego pilota. Jednakże, choć ten etap może być atrakcyjny, jego realizacja napotyka kilka problemów technicznych i koncepcyjnych: (1) Reaktywność: system musi reagować na działania użytkownika w odpowiednio krótkim czasie. W

przeciwnym razie użytkownik jest rozproszony, sfrustrowany i zagubiony w odniesieniu do stanu komunikacyjnego. Opracowano odpowiednie techniki rozpoznawania twarzy, wykrywania spojrzeń i rozpoznawania gestów, a także wyznaczają obszar aktywnych badań . (2) Solidność: Wysokie współczynniki wykrywalności wyników fałszywie dodatnich mogą skutkować niepożądanym przez użytkownika zachowaniem systemu i sprzecznym z jego oczekiwaniami. Jest to szczególnie problematyczne, ponieważ nie wszystkie zachowania użytkowników są kierowane do systemu. W tym przypadku ważną koncepcją jest wspólna uwaga – stan, w którym obaj partnerzy komunikacji skupiają się na tej samej rzeczy i są świadomi wzajemnej uwagi . Na przykład w interakcji człowiek-robot robot musi wykryć, kiedy użytkownik jest skierowany w jego stronę. Jednocześnie głowa i oczy robota będą śledzić twarz użytkownika, aby wzmocnić nawiązaną komunikację. (3) Niezawodność: Działania użytkownika częściowo przeoczone przez system mogą spowodować uszkodzenie wszystkich danych wprowadzanych przez użytkownika do systemu. Zatem musi istnieć sposób określenia, czy dane wejściowe są dobrze uformowane, czy nie. Jest to trudny problem uczenia się i rozpoznawania, ponieważ ludzie zazwyczaj wykonują zadania o dużej zmienności i nie są świadomi ograniczeń systemu. Ciekawym kierunkiem badań jest zrozumienie, w jaki sposób ludzie komunikują oczekiwania w dialogu, na przykład zadając pytanie tak/nie lub stosując inne konwencje, które ograniczają możliwe odpowiedzi. (4) Sytuacyjność: Interpretacja większości ludzkich zachowań zależy od kontekstu. Dlatego wiele systemów zaprojektowano dla bardzo konkretnego scenariusza lub domeny aplikacji. Aby pokonać te ograniczenia, ważnym pojęciem jest świadomość kontekstu – koncepcja wprowadzona w społeczności komputerów mobilnych . W przypadku widzenia komputerowego został on zastosowany na przykład w pomieszczeniach percepcyjnych przez Crowleya i innych (2002). Tam działania człowieka są obserwowane za pomocą wielu kamer i kategoryzowane ze względu na różne konteksty i sytuacje. W konsekwencji punktów omówionych powyżej badania nad interakcją człowiek-maszyna opartą na wizji zawsze muszą uwzględniać kompletne systemy wraz z ich partnerami interakcji, co czyni je zadaniem wysoce interdyscyplinarnym. Większość systemów w tym obszarze ściśle ogranicza ustawienia komunikacyjne. Wczesne prace zostały wykonane przez Bolta i jego współpracowników (Bolt 1980) nad jego systemem „Put-That-There". Użytkownik mógł tworzyć i przesuwać elementy geometryczne na ekranie za pomocą gestów i poleceń głosowych. Dzisiejsze systemy obejmują szerokie spektrum technik i zastosowań. SafetyEYE opracowany w ramach badań branżowych szacuje promień działania przemysłowego robota produkcyjnego i zatrzymuje go w przypadku ingerencji człowieka i maszyny. MIT Kidsroom zapewnia dzieciom interaktywną przestrzeń do zabawy narracyjnej . Opiera się na technikach wizualnego rozpoznawania działań, które są połączone z kontrolą obrazów, wideo, światła, muzyki, dźwięku i narracji. Crowley i inni opisują interaktywną Magiczną Tablicę opartą na śledzeniu palców i oknie percepcyjnym, które przewija się poprzez wykrywanie ruchów głowy. W ostatnich latach śledzenie ciała stało się gorącym tematem komercyjnym w przypadku konsol do gier, takich jak PlayStation firmy Sony i Xbox firmy Microsoft. Inaczej postawiono w systemie VAMPIRE który pomagał ludziom w codziennych zadaniach, prowadząc ich krok po kroku przez przepis. Zostało to zademonstrowane w scenariuszu mieszania napojów i wykorzystano techniki rozpoznawania obiektów, śledzenia, lokalizacji i rozpoznawania działań w celu uzyskania pomocy dla użytkownika w oparciu o techniki rzeczywistości rozszerzonej. Wiele pracy włożono w to, aby wypełnić lukę w komunikacji między ludźmi a robotami usługowymi, których zadaniem jest pełnienie roli towarzysza w domu. Przykładami są PR2 firmy Willow Garage, Care-O-Bot 3 firmy Fraunhofer IPA, Cosero z Uniwersytetu w Bonn lub ToBI z Bielefeld. W pierwszym z nich możesz złożyć pranie lub wypić napój z lodówki. Pozostali brali aktywny udział w konkursie RoboCup@Home, który obejmuje szereg testów porównawczych, począwszy od śledzenia osób i przedstawiania im gości, po sprzątanie i przynoszenie napojów. W porównaniu z komunikacją człowiek-człowiek (HHC) interakcja człowiek-maszyna jest nadal krucha i znajduje się w powijakach. Dzisiejsze badania koncentrują się na naśladowaniu pewnych aspektów HHC w celu stawienia czoła czterem opisanym wyzwaniom.

Kontekstowe zrozumienie sceny

Większość podejść do widzenia komputerowego nie interpretuje całych obrazów, ale wybrane ich części. Ich celem jest wyodrębnienie obiektów pierwszego planu z bałaganu w tle. Następnie każdy obiekt jest klasyfikowany oddzielnie. Tło jest ignorowane i postrzegane jako nieistotne, rozpraszające dane lub po prostu jako szum. Kontekstowe zrozumienie sceny opiera się na kontrastującym założeniu, że obiekty na pierwszym planie nie mogą zostać automatycznie wyodrębnione lub przynajmniej nie dostarczają wystarczających informacji do klasyfikacji. Przetwarza zignorowane wcześniej dane – bałagan w tle i informacje relacyjne – w celu wyciągnięcia możliwych interpretacji dla obiektów na pierwszym planie. Celem tych technik jest zatem włączenie kontekstu sceny do procesu klasyfikacji. Pionierskie prace przeprowadzili Strat i Fischler , którzy zdefiniowali zbiory kontekstów rządzące wywoływaniem etapów przetwarzania systemu. Identyfikują cztery różne rodzaje kryteriów składających się na zbiory kontekstów: (1) konteksty globalne – atrybuty całej sceny, takie jak dzień lub krajobraz; (2) lokalizacja – konfiguracja przestrzenna sceny, np. dotknięcie ziemi lub zbieżność z innymi typami obiektów; (3) wygląd sąsiadujących obiektów, np. podobieństwo lewego i prawego oka twarzy; oraz (4) funkcjonalność – rola obiektu w scenie, np. wspieranie innego obiektu lub mostkowanie strumienia. Z kontrolnego punktu widzenia Strat i Fischler stosują trzy rodzaje operacji opartych na kontekście, aby kierować procesem interpretacji sceny: generowanie hipotez, walidacja hipotez i porządkowanie hipotez. W trakcie poszukiwania hipotez (generowania) konstruowane są spójne grupy rozpoznanych bytów, które reprezentują częściowe interpretacje danej sceny. Główną wadą tego rodzaju podejścia jest ogromne zadanie inżynierii wiedzy polegające na kodowaniu wiedzy kontekstowej systemu. Jednakże ogólne typy wprowadzonych kontekstów i różne rodzaje zaprojektowanych zasad sterowania są nadal aktualne w obecnym stanie techniki. W późniejszych pracach zaadaptowano modele probabilistyczne do interpretacji kontekstowej, które w systematyczny sposób wychwytują relacje i niepewność. Poniższe przykłady ilustrują nowsze trendy w odniesieniu do wprowadzonych wcześniej typów kontekstu ogólnego. Konteksty globalne służą do klasyfikacji miejsc semantycznych (np. ulica, miasto, plaża lub kategorie pomieszczeń wewnętrznych, takich jak kuchnia). W ten sposób obliczana jest holistyczna reprezentacja obrazu – tzw. istota obrazu. Kategoria semantyczna określa oczekiwania dotyczące często występujących obiektów (takich jak te zwykle spotykane w kuchni). Lokalizacja jest modelowana przez Hoiema i współpracowników , którzy wiążą wykrycie obiektów z ogólnym kontekstem sceny 3D oraz oceniają skalę i lokalizację w odniesieniu do szacowanej geometrii sceny. Funkcjonalność jest wykorzystywana przez Moore'a, Essę i Hayesa którzy wiążą ludzkie działania i przedmioty za pomocą modelu probabilistycznego. Wprowadzają koncepcję przestrzeni obiektowych, które łączą oba rodzaje informacji w przestrzeni i czasie. Wreszcie konteksty językowe odnoszą się do dodatkowych informacji podanych przez tekst równoległy lub mowę. Tego rodzaju dane bimodalne często pojawiają się w katalogach, gazetach, czasopismach, stronach internetowych, wiadomościach telewizyjnych, filmach lub dialogach interakcji człowiek-maszyna. Informacja werbalna obejmuje zasadniczo wszystkie trzy typy informacji kontekstowych. Podpis obrazu zawierający wzmiankę o „Nowym Jorku" lub „ruchu ulicznym" może wskazywać, że zdjęcie przedstawia scenę miejską. Inne opisy werbalne, na przykład dwie osoby stojące obok siebie, zapewniają lokalne ograniczenia dla analizy obrazu. Konteksty funkcjonalne można wyprowadzić z czasowników, choć metoda ta nie jest powszechnie stosowana.

Podsumowanie i wnioski

Agenci, ludzcy lub sztuczni, muszą postrzegać swoje środowisko, aby w nim działać i przetrwać. Percepcja wzrokowa jest najsilniejszym zmysłem człowieka, a praca w dziedzinie widzenia komputerowego ma na celu zapewnienie wymaganych możliwości. Podsumowano główne osiągnięcia, zaczynając od przeglądu trendów i perspektyw, a następnie podkreślając główne obszary zastosowań.

Obecnie maszyny mogą uczyć się, a następnie rozpoznawać obiekty na podstawie obrazów 2D zawierających do około 1000 obiektów, a liczba ta stale rośnie. Jest on jednak ograniczony do baz danych obrazów, w których rozmiary obiektów lub typowych scen są podobne. W otwartych środowiskach, takich jak poszukiwania w domach, różnice w oświetleniu, punkcie widzenia lub okluzji nadal stanowią wyzwanie. Korzystając z obrazów 3D, na przykład skanerów laserowych lub obrazów głębi, można uzyskać kształt obiektów i wykorzystać go do sterowania procesami przemysłowymi, takimi jak chwytanie przez roboty lub malowanie natryskowe. Śledzenie obiektów lub punktów szczególnych w dłuższych sekwencjach wideo można przeprowadzić w czasie rzeczywistym, jeśli zapewniona jest wystarczająca tekstura. Ustanowiono zasady wykorzystania informacji o obrazie oraz skutecznego przewidywania i wyszukiwania kolejnych obrazów, a także dostępne są wizualne metody obsługi ramion robotów. Wydajność i niezawodność w czasie rzeczywistym osiągane przez dzisiejsze techniki widzenia komputerowego do śledzenia dłoni, śledzenia ciała ludzkiego, rozpoznawania twarzy itd. prowadzą do nowej jakości interakcji człowiek-maszyna opartej na wizji. Omówiliśmy kilka wyzwań w tej nowej dziedzinie, która łączy obszary widzenia komputerowego (CV) i interakcji człowiek-komputer (HCI). W ciągu ostatnich lat powstało kilka nowych serii warsztatów, takich jak CV4HCI i CV skoncentrowane na człowieku. Oczekujemy, że to małżeństwo zapewni dalsze owocne wpływy na boisku, przyjmując dwie perspektywy: jak projektować systemy CV dla użytkowników i jak skutecznie włączyć użytkownika w pętlę przetwarzania wizualnego. Jednym z wyzwań było umiejscowienie: biorąc pod uwagę dowolną sytuację podczas interakcji, kiedy i jakimi informacjami należy zaniepokoić użytkownika? To samo pytanie można zadać w odniesieniu do systemu wizyjnego. Nie wszystkie informacje są ważne i nie wszystkie wyniki wykrywania są ważne. Pojęcie kontekstu z jednej strony dostarcza pojęcia globalnej spójności, a z drugiej strony ramy znaczeniowej. Nawet przy dość wyrafinowanych i wydajnych technikach rozpoznawania kontekst zachowa swoją rolę, gdy mówimy o komputerowych systemach wizyjnych, które muszą działać w rzeczywistych środowiskach. Komputerowe systemy wizyjne muszą łączyć techniki do celów aplikacyjnych. To jest rdzeń CV jako dyscypliny inżynierskiej. Jednakże na przestrzeni lat udowodniono, że trudno jest zdefiniować ogólną architekturę, która integruje wszystkie komponenty potrzebne do różnych zastosowań. Niektóre podejścia wykazały swoje zastosowanie w udanych projektach europejskich obejmujących wielu partnerów . Prawdziwy postęp jest trudny do osiągnięcia od strony teoretycznej i wymaga potwierdzenia przez praktyczną realizację systemów . Chociaż wyniki te wskazują, jak bardzo rozwinęła się ta dziedzina, przed nią stoi kilka wyzwań. Na przykład prace nad rozpoznawaniem klas obiektów ograniczają się obecnie do kilku najistotniejszych klas, takich jak koła czy samoloty; należy rozszerzyć możliwość wykrywania punktów uchwytu na dowolnych obiektach z lokalizacji obiektów płaskich do lokalizacji obiektów w pełnym 3D; i nie jest jeszcze możliwe wywnioskowanie funkcji przedmiotu z obrazowania jego kształtu. Niemniej jednak istnieje nadzieja, że wizja komputerowa będzie w coraz większym stopniu integrowana z innymi sztuczną inteligencjąmetody budowania bardziej kompletnych systemów.