



دانشکده مهندسی کامپیوتر

ردیابی جهت نگاه انسان با استفاده از مدل‌های گرافیکی احتمالاتی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته مهندسی کامپیوتر گرایش هوش مصنوعی

نام دانشجو

رحیم انتظاری

اساتید راهنما:

دکتر محمود فتحی

دکتر رضا برنگی

۱۳۹۵ بهمن‌ماه

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

تأییدیهی هیأت داوران جلسه‌ی دفاع از پایان‌نامه/رساله

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: رحیم انتظاری

عنوان پایان‌نامه یا رساله: ردیابی جهت نگاه انسان با استفاده از مدل‌های گرافیکی احتمالاتی

تاریخ دفاع: ۱۳۹۵/۱۱/۲۵

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما	محمود فتحی	استاد	علم و صنعت ایران	
۲	استاد راهنما	رضا برنتگی	دانشیار	علم و صنعت ایران	
۶	استاد مدعو خارجی	علی ذاکر الحسینی	استادیار	شهید بهشتی	
۸	استاد مدعو داخلی	محسن سریانی	دانشیار	علم و صنعت ایران	

تأییدیه‌ی صحت و اصالت نتایج

با اسمه تعالی

اینجانب رحیم انتظاری به شماره دانشجویی ۹۳۷۲۲۰۲۵ دانشجوی رشتهمهندسی کامپیوتر گرایش هوش مصنوعی..... مقطع تحصیلی...کارشناسی ارشد.... تأیید می‌نمایم که کلیه‌ی نتایج این پایان‌نامه/رساله حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انصباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احراق حقوق مکتب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: رحیم انتظاری

امضا و تاریخ:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنمای شرح زیر تعیین می‌شود، بلامانع است:

- بهره‌برداری از این پایان‌نامه/رساله برای همگان بلامانع است.
- بهره‌برداری از این پایان‌نامه/رساله با اخذ مجوز از استاد راهنمای، بلامانع است.
- بهره‌برداری از این پایان‌نامه/رساله تا تاریخ ممنوع است.

نام استاد یا استادید راهنمای:

تاریخ:

امضا:

تقدیم به:

مقدس‌ترین واژه‌ها ، پدر و مادر مهربانم که زندگیم را مديون مهر و عطوفت آن‌ها می‌دانم
همسرم که نشانه لطف الهی در زندگی من است
و برادرانم، همراهان همیشگی زندگیم

تشکر و قدردانی:

از استادان گرامی؛ جناب آقایان دکتر فتحی و دکتر برنگی که در کمال سعه صدر، با حسن خلق و فروتنی،
از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند؛

و از استادان گرامی؛ جناب آقایان دکتر ذاکرالحسینی و دکتر سریانی که زحمت داوری این رساله را متقبل
شدند؛ کمال تشکر و قدردانی را دارم.

چکیده

تخمین جهت نگاه انسان شامل تشخیص و ردیابی چشم، ردیابی حرکات آن و ارائه مدل محاسباتی برای تخمین جهت نگاه انسان است. تشخیص چشم و تخمین جهت نگاه انسان نقش مهمی در بیان خواسته‌های فرد، نیازها، فرآیندهای شناختی، حالات احساسی و عاطفی و ارتباطات بین افراد ایفا می‌کند. ویژگی‌های هندسی، نورسنجی و حرکتی چشم‌ها سرنخ‌های دیداری مهمی برای تشخیص و بازشناسی صورت و درک حالات صورت به دست می‌دهند. هم‌چنین تخمین جهت نگاه کاربردهای فراوانی از جمله تحلیل توجه انسان، رابطه‌های کاربری مبتنی بر چشم و تشخیص فعالیت انسان دارد.

روش‌های مختلفی برای مدل‌سازی ردیابی چشم و تخمین جهت نگاه به عنوان یکی از مهم‌ترین اجزای صورت در تصاویر در سال‌های اخیر ارائه و نتایج قابل توجهی در کارهای مختلف مطرح شده‌اند. بدین منظور مدل‌های ارائه شده در این زمینه را بررسی کرده، به مقایسه و ارزیابی آن‌ها می‌پردازیم. از آن‌جا که در سال‌های اخیر پیشرفت‌های شگرفی در حوزه‌های مختلف هوش مصنوعی به کمک یادگیری عمیق حاصل شده است در این پژوهش نیز از یادگیری عمیق به منظور استخراج ویژگی استفاده شده است. در طرف دیگر مدل‌های گرافیکی احتمالاتی قدمت طولانی‌تری در کشف ساختار موجود در مساله داشته‌اند. در این پژوهش معماری جدیدی بر اساس تلفیق شبکه‌های عصبی پیچشی عمیق و هم‌چنین مدل‌های گرافیکی احتمالاتی ارائه شده است. در معماری مذکور ویژگی‌های مورد نیاز برای آموزش مدل گرافیکی به کمک یادگیری عمیق استخراج شده و در اختیار میدان‌های تصادفی شرطی به عنوان یکی از معروف‌ترین مدل‌های گرافیکی احتمالاتی قرار می‌گیرند. به منظور نمایش دقیق معماری پیشنهادی، مجموعه دادگان EYEDIAP و MPIIGaze انتخاب و چند آزمایش مختلف از جمله معماری پیشنهادی بر روی مجموعه دادگان مذکور مورد بررسی و مقایسه قرار گرفته‌اند. میانگین خطای تخمین جهت نگاه در این دو مجموعه دادگان برای معماری پیشنهادی به ترتیب ۷,۵ و ۶,۴ درجه بوده است.

واژه‌های کلیدی :

ردیابی چشم، تخمین جهت نگاه، شبکه‌های عصبی عمیق، شبکه‌های پیچشی، مدل‌های گرافیکی احتمالاتی

فهرست مطالب

۱	فصل ۱: مقدمه
۲	۱-۱- مقدمه
۳	۲-۱- بیان مسئله پژوهش
۳	۳-۱- اهمیت انجام پژوهش
۴	۴-۱- اهداف پژوهش
۵	۵-۱- امکانات و محدودیت‌های پژوهش
۵	۶-۱- ساختار پایان‌نامه
۶	فصل ۲: مروری بر منابع
۷	۱-۲- مقدمه
۷	۲-۲- مروری بر ادبیات موضوع
۹	۲-۲-۲- رویکردهای مبتنی بر شکل
۱۵	۳-۲-۲- مدل‌های ترکیبی
۱۶	۳-۲- سایر روش‌ها
۱۸	۴-۲- کارهای مرتبط جدید
۱۹	۱-۴-۲- تشخیص جهت نگاه از روی ظاهر در محیط طبیعی
۱۹	۲-۴-۲- ردیابی چشم برای همه
۲۱	۳-۴-۲- تشخیص جهت نگاه از روی تصویر صورت
۲۳	۵-۲- نتیجه‌گیری
۲۴	فصل ۳: مبانی تحقیق
۲۵	۱-۳- مقدمه
۲۵	۲-۳- شبکه‌های عصبی
۲۶	۲-۲-۳- شبکه‌های عصبی سنتی
۲۶	۳-۲-۳- شبکه‌های عصبی عمیق
۴۰	۳-۳- مدل‌های گرافیکی احتمالاتی
۵۲	۲-۳-۳- علت انتخاب روش
۵۳	فصل ۴: روش تحقیق
۵۴	۱-۴- مقدمه
۵۴	۲-۴- جزئیات پیاده‌سازی
۵۶	۲-۲-۴- روش‌های بهینه سازی
۶۰	۳-۲-۴- آزمایش‌ها
۶۵	۳-۴- معماری سامانه

۶۶	۲-۳-۴- معماری بخش یادگیری عمیق
۷۱	۳-۳-۴- معماری بخش مدل گرافیکی احتمالاتی
۷۶	۴-۴- نتیجه‌گیری
۷۷	فصل ۵: نتایج و تفسیر آن‌ها
۷۸	۱-۵- مقدمه
۷۸	۲-۵- نتایج
۸۲	فصل ۶: جمع‌بندی و پیشنهادها
۸۳	۱-۶- جمع‌بندی
۸۳	۲-۶- پیشنهادها
۸۵	مراجع
۹۱	پیوست‌ها

فهرست اشکال

..... ۸ شکل (۱-۲) تغییرات شکل چشم از زوایای مختلف
..... ۱۷ شکل (۲-۲) مدل گرافیکی استفاده شده در [۲۶]
..... ۱۸ شکل (۳-۲) گراف فاکتور استفاده شده در [۲۸]
..... ۱۹ شکل (۴-۲) معماری شبکه عمیق استفاده شده در [۲۹]
..... ۲۰ شکل (۵-۲) معماری شبکه عمیق استفاده شده در [۳۰]
..... ۲۲ شکل (۶-۲) معماری شبکه عمیق پیشنهادی در [۳۱]
..... ۲۲ شکل (۷-۲) مقایسه عملکرد معماری پیشنهادی [۳۱]
..... ۲۵ شکل (۱-۳) نمونه‌ای از ساختار یک پرسپترون ساده [۳۲]
..... ۲۸ شکل (۲-۳) نموداری کلی از حوزه‌های یادگیری عمیق
..... ۲۹ شکل (۳-۳) بازنمایی ویژگی‌های استخراج شده در لایه‌های مختلف شبکه‌های عمیق [۳۷]
..... ۳۰ شکل (۴-۳) ساختار معماري خود رمزکننده با یک لایه مخفی [۳۸]
..... ۳۳ شکل (۵-۳) نمونه‌ای از لایه پیچش و ادغام بر روی داده ورودی [۴۰]
..... ۳۴ شکل (۶-۳) نمونه‌ای از مراحل عملیات پیچش بر روی داده ورودی [۴۱]
..... ۳۵ شکل (۷-۳) خروجی نهایی نمونه ورودی پیچشی در [۴۱]
..... ۳۶ شکل (۸-۳) نمونه‌ای از عملیات ادغام ویژگی‌های پیچشی [۴۲]
..... ۳۷ شکل (۹-۳) خروجی تابع فعال‌ساز سیگموئید
..... ۳۸ شکل (۱۰-۳) تابع فعال‌ساز Tanh
..... ۳۹ شکل (۱۱-۳) خروجی تابع فعال‌ساز ReLU
..... ۴۰ شکل (۱۲-۳) تصویری از مقاله [۴۳] برای نمایش بهبود شش برابری همگرایی با ReLU (خط ساده) در برابر همگرایی با Tanh (خط‌چین)
..... ۴۱ شکل (۱۳-۳) مدل گرافیکی احتمالاتی از دو بخش نظریه احتمال و گراف تشکیل شده است
..... ۴۴ شکل (۱۴-۳) یک مثال از دسته‌بند بیز ساده
..... ۴۴ شکل (۱۵-۳) مثالی از یک گراف جهت‌دار و توزیع احتمال مربوط به آن
..... ۵۵ شکل (۱-۴) مقایسه سرعت همگرایی روش‌های بهینه‌سازی مختلف [۴۸]
..... ۶۰ شکل (۲-۴) شبکه LeNet-5
..... ۶۱ شکل (۳-۴) شبکه LeNet تغییریافته در آزمایش شماره ۱
..... ۶۵ شکل (۴-۴) معماری شبکه VGG-16 [۵۸]
..... ۶۶ شکل (۵-۴) معماری سامانه پیشنهادی
..... ۶۷ شکل (۶-۴) معماری شبکه AlexNet [43]
..... ۶۹ شکل (۷-۴) معماری بخش یادگیری عمیق از سامانه پیشنهادی
..... ۶۹ شکل (۸-۴) شبکه پیچشی استفاده در زیر بخش یادگیری عمیق

شکل (۹-۴) جزئیات شبکه عمیق استفاده شده در سامانه پیشنهادی.....	۷۰
شکل (۱۰-۴) یک مثال از خروجی شبکه عمیق استفاده شده در سامانه پیشنهادی.....	۷۱
شکل (۱۱-۴) نمایی ساده از میدان‌های تصادفی شرطی (سمت چپ)- میدان‌های تصادفی شرطی نهان(سمت راست).....	۷۳
شکل (۱۲-۴) معماری زیر بخش مدل‌های گرافیکی احتمالاتی در سامانه پیشنهادی.....	۷۴
شکل (۱۳-۴) بخش‌بندی نمایشگر استفاده شده در مجموعه دادگان.....	۷۵
شکل (۱-۵) مقایسه دقต روشهای مختلف بر روی مجموعه دادگان EYEDIAP	۷۸
شکل (۲-۵) مقایسه دقت روشهای مختلف بر روی مجموعه دادگان MPIIGaze	۷۹
شکل (۱-۶) نمونه‌هایی از فریم‌های ضبط شده در مجموعه دادگان EYEDIAP [۶۸]	۹۳
شکل (۲-۶) نمونه‌هایی از تصاویر موجود در مجموعه دادگان MPIIGaze [۲۹]	۹۴
شکل (۳-۶) برخی مشخصات مجموعه دادگان MPIIGaze [۲۹]	۹۴

فهرست جداول

- جدول (۱-۲) جدول مقایسه‌ای تأثیر حذف زیر بخش‌های معماری پیشنهادی [۳۰] ۲۱
جدول (۱-۳) نمادهای به کار رفته در خودرمزنده ۳۲
جدول (۱-۴) جزئیات لایه‌های مختلف AlexNet ۶۹
جدول (۱-۶) خلاصه وضعیت جلسات ضبط ویدئو در مجموعه دادگان [۶۸] ۹۳

فصل ۱:

مقدمه

۱-۱- مقدمه

مهم‌ترین راه‌های برقراری ارتباط غیرکلامی شامل حالات صورت، تکان دادن دست‌ها و جهت نگاه می‌شود. در این میان جهت نگاه انسان نقش مهمی در بیان خواسته‌های فرد، نیازها، فرآیندهای شناختی، حالات احساسی و عاطفی و ارتباطات بین افراد ایفا می‌کند. به لحاظ فنی تشخیص جهت نگاه انسان یکی از موضوعات تحقیقاتی چالش‌برانگیز در حوزه بینایی ماشین و پردازش تصویر است.

در گذشته و به خصوص سال‌های اخیر تحقیقات فراوانی در زمینه‌ی ردیابی چشم و تخمین جهت نگاه انسان انجام‌شده است. در حالی که روش‌های مبتنی بر رگرسیون ساده و نسبتاً دقیق هستند اما به دلیل حساسیتشان به حرکات سر فقط مناسب برخی کاربرها هستند. سامانه‌های مبتنی بر مدل سه بعدی می‌توانند با حرکات طبیعی سر کنار بیایند اما آن‌ها معمولاً به درجه‌بندی هندسی^۱ نیازمندند. افزودن یک دوربین دیگر ممکن است تعداد نقاط درجه‌بندی را کم کند اما از طرفی موجب افزایش هزینه می‌شود. در طرف دیگر روش‌های مبتنی بر ظاهر خیلی به شرایط راه‌اندازی و با نوری وابسته نیستند و بنابراین ساده‌تر و منعطف‌تر هستند؛ اما این روش‌ها نیاز به درجه‌بندی بیش‌تری داشته و عدم وابستگی به حرکات سر را تضمین نمی‌کنند. علیرغم تحقیقات فراوان کماکان تخمین جهت نگاه انسان به عنوان یک مبحث چالش‌برانگیز مطرح می‌شود.

به‌طور خلاصه سامانه‌های ردیابی جهت نگاه در آینده می‌بایست کم‌هزینه بوده، به راحتی نصب شده، کم‌ترین درجه‌بندی را داشته و در شرایط نوری متغیر و با حرکات طبیعی سر دقت قابل قبولی داشته باشند. در پژوهش پیش‌رو تلاش شده است به بررسی نیازمندی‌ها و کاستی‌های تشخیص جهت نگاه انسان پرداخته شود. تشخیص نگاه انسان کاربردهای فراوانی داشته و به‌طورکلی با در اختیار داشتن جهت نگاه انسان می‌توان اطلاعات مهمی درباره هدف و میزان توجه فرد به دست آورد.

در این فصل به بیان دقیق مسئله پژوهش، اهمیت انجام پژوهش، امکانات و محدودیت‌های انجام این پژوهش و معرفی سازمان پایان‌نامه پرداخته شده است.

¹ Geometric calibration

۱-۲- بیان مسئله پژوهش

همان‌طور که در عنوان این پژوهش ذکر شده است، به دنبال ارائه سامانه‌ای مبتنی بر یادگیری عمیق و مدل‌های گرافیکی احتمالی هستیم که توانایی تشخیص جهت نگاه انسان را دارد. گام‌های مختلفی برای این مسئله قابل تعریف است که از جمله آن‌ها می‌توان به موارد زیر اشاره کرد:

- پیش‌پردازش‌های لازم بر روی ویدئو/تصویر دریافتی
- طراحی و ارائه معماری موردنیاز بهمنظور استخراج ویژگی از فریم‌های موجود
- طراحی و ارائه مدلی برای آموزش دسته‌بندی ویژگی‌های موجود در فریم‌ها
- پردازش‌های موردنیاز بهمنظور ارائه خروجی نهایی سامانه

تعاریف متعددی برای جهت نگاه انسان در ادبیات مطرح شده است؛ اما پرکاربردترین تعریف از جهت نگاه انسان به نقطه‌ای یاد می‌کند که انسان به آن می‌نگرد.^۱ در این پژوهش نیز منظور از تخمین جهت نگاه انسان تشخیص نقطه‌ای است که کاربر به آن نگاه می‌کند. پس از تشخیص این نقطه می‌توان برداری را تعریف کرد که مرکز چشم را به نقطه تشخیصی متصل می‌کند. بردار دیگری نیز از اتصال مرکز چشم به نقطه‌ای که بیننده واقعاً به آن نگاه می‌کند (برچسب موجود در مجموعه دادگان) قابل تعریف است. بدین‌ترتیب منظور از خط، فاصله بین این دو بردار بر حسب درجه است.

۱-۳- اهمیت انجام پژوهش

همان‌طور که بیان شد جهت نگاه انسان نقش مهمی در بیان خواسته‌های فرد، نیازها، فرآیندهای شناختی، حالات احساسی و عاطفی و ارتباطات بین افراد ایفا می‌کند. به دلیل این اهمیت و با پیشرفت روزافرون فناوری، تشخیص جهت نگاه انسان کاربردهای فراوانی را به خود اختصاص داده است. خواص حرکت چشم موجب شده است سامانه‌های ردیابی جهت نگاه به ابزاری منحصر به فرد و مناسب برای افراد معلول تبدیل شوند تا برایشان حرکت چشم راهی شود برای برقراری ارتباط و تعامل با سایر افراد و رایانه. برای مثال در بین کاربردها می‌توان به نوشتن با کمک چشم اشاره کرد که کاربر از طریق ورودی‌های جهت نگاه متن موردنظر خود را تولید می‌کند. ردیابی چشم در صنعت خودرو نیز مورد توجه واقع شده است. برای مثال در این زمینه می‌توان بر میزان هوشیاری راننده نظارت داشت [۱]. مطالعه بر روی تشخیص و ردیابی

¹ Point of Regard

جهت نگاه همچنین می‌تواند در نصب دوربین‌هایی در خودروها به کار رود تا بهوسیله آن‌ها رفتار دیداری راننده خودرو به‌طور زمان واقع^۱ ارزیابی شود. از انعکاس قرنیه^۲ برای آزمون جراحان نیز استفاده شده است [۲] تا میزان زمانی که آن‌ها به ابزارآلات جراحی توجه می‌کنند سنجیده شود. در علم روان‌شناسی ردیابی چشم و جهت نگاه به دانشمندان کمک می‌کند تا به بررسی روند رشد و تغییر ادراک شناخت و توانایی‌های اجتماعی از دوران کودکی تا بزرگسالی بپردازند.

تشخیص و ردیابی جهت نگاه همچنین می‌تواند در ارزیابی رفتار مشتری بر اساس میزان جلب توجه آن‌ها در هنگام خرید به طراحی بسته بندی و مکان قرارگیری محصول در فروشگاه به کار رود. همچنین می‌توان از جهت نگاه در ساخت کلیپ‌های تبلیغاتی استفاده کرد تا درک کنیم چه مناطقی از تصاویر تبلیغاتی بیشتر مورد توجه مشتری واقع می‌شود. در علوم ورزشی نیز به دلیلی آن‌که هماهنگی چشم و عضلات از جمله دست بسیار مهم است ردیابی چشم می‌تواند به دانشمندان این حوزه و همچنین بازآموزان ورزشی کمک کند تا پیشرفت خود را در زمینه ورزشی مورد نظر مطالعه کنند [۳].

۴-۱- اهداف پژوهش

هدف از این پژوهش طراحی سامانه‌ای است که توانایی تشخیص جهت نگاه انسان را داشته باشد؛ بدین منظور در این پژوهش تلاش می‌شود تا سامانه‌ای عمومی با کمترین میزان محدودیت برای کاربر به صورت ابتدا به انتهای طراحی شود.

با توجه به استفاده از روش یادگیری عمیق که یکی از مباحث جدید در حوزه یادگیری ماشین است، انتظار می‌رود ویژگی‌های استخراج شده توسط شبکه‌های عصبی پیچشی^۳ بسیار مفیدتر از ویژگی‌هایی باشند که به صورت دستی^۴ استخراج شده‌اند، به این معنی که ویژگی‌های استخراج شده به کمک یادگیری عمیق توانایی بیشتری در تفکیک داشته باشند. از طرف دیگر انتظار می‌رود استفاده از مدل‌های گرافیکی احتمالاتی که یکی از قوی‌ترین روش‌های مدل‌سازی در یادگیری ماشین محسوب می‌شوند بتواند به خوبی از ویژگی‌های استخراج شده به منظور تخمین جهت نگاه استفاده کند.

¹ Real time

² Cornea Reflection

³ Convolutional neural networks

⁴ Hand craft features

۱-۵- امکانات و محدودیت‌های پژوهش

در این پژوهش، امکانات سخت‌افزاری و نرم‌افزاری مختلفی مورد نیاز است. در بخش نرم‌افزاری مجموعه دادگان‌های مختلفی وجود دارند. در این پژوهش از دو مورد از این مجموعه‌های دادگان استفاده شده است که به نسبت سایر مجموعه‌های دادگان برتری داشته‌اند. سایر نیازمندی‌های نرم‌افزاری شامل سامانه‌عامل اوبونتو ۱۴.۰۴، نرم‌افزار MATLAB 2016، چارچوب یادگیری عمیق Caffe است.

امکانات سخت‌افزاری مورد نیاز این پژوهش، رایانه مناسب جهت پردازش‌های مبتنی بر شبکه‌های عصبی عمیق است. برای بخش آموزش شبکه‌های عصبی به رایانه‌ای با پردازنده اینتل سری Core i5، حافظه اصلی ۸ گیگابایتی و پردازنده گرافیکی Nvidia، پشتیبانی‌کننده از قابلیت CUDA نیاز است. در این پژوهش از پردازنده گرافیکی GTX 980 Ti استفاده شده است.

۱-۶- ساختار پایان‌نامه

در فصل نخست مقدمه‌ای در حوزه تشخیص جهت نگاه انسان، کاربردها و اهمیت انجام این پایان‌نامه بیان شد. در فصل دوم مروری بر کارهای انجام‌شده توسط سایر محققان در رדיابی چشم و تشخیص جهت انسان خواهیم داشت. آزمایش‌های مختلف انجام‌شده، مفاهیم استفاده شده در سامانه نهایی در فصل سوم با عنوان مبانی تحقیق توضیح داده شده است. در فصل چهارم، روش تحقیق، سامانه نهایی به همراه جزئیات پیاده‌سازی بیان شده است. نتایج به دست آمده از ارزیابی آزمایش‌های مطرح شده به همراه نتایج دقت سامانه پیشنهادی در فصل پنجم مطرح شده‌اند. در فصل ششم نیز نتیجه‌گیری و پیشنهادهای آینده بیان شده‌اند. فصل پیوست شامل بررسی مجموعه‌های دادگان استفاده شده در این پایان‌نامه می‌شود. در انتهای نیز مراجع مورداستفاده در این پژوهش ذکر شده‌اند.

فصل ۲:

مروری بر منابع

۱-۲ - مقدمه

تخمین جهت نگاه انسان یکی از مسائل چالش‌برانگیز در حوزه بینایی ماشین محسوب می‌شود. این تخمین می‌تواند با تشخیص و ردیابی چشم همراه باشد و یا به‌طور مستقیم انجام شود. در زمینه‌ی تشخیص چشم مهم است مدلی از چشم به دست آید به‌طوری که این مدل هم به‌اندازه کافی جامع باشد تا بتواند تغییرات عمدی در ظاهر و پویایی را پوشش داده و هم استفاده از آن از نظر محاسباتی مقرن به‌صرفه باشد. ظاهر چشم مشترکات زیادی در بین نژادهای مختلف شرایط نوری و زاویه دید دارد اما حتی برای یک شی یکسان تغییر کوچکی در زاویه دید ممکن است منجر به تغییرات عمدی در ظاهر چشم شود.

علیرغم تحقیقات فراوان کماکان به دلایلی همچون انسداد^۱ چشم توسط پلک‌ها، بسته و یا باز بودن چشم‌ها، تغییرات در مقیاس و یا حرکت سر شناسایی و ردیابی چشم و پساز آن تخمین جهت نگاه انسان به عنوان یک مبحث چالش‌برانگیز مطرح می‌شود.

۲-۲ - مروری بر ادبیات موضوع

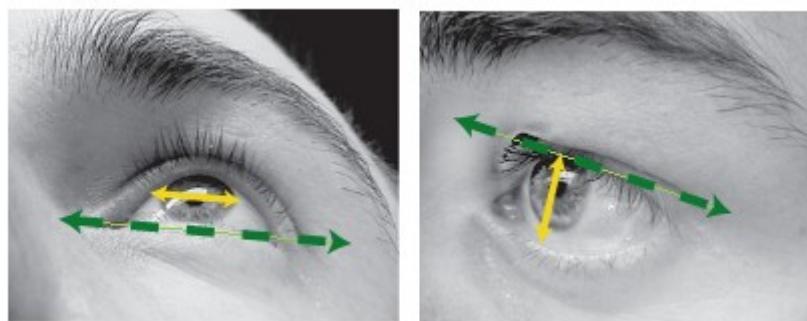
علیرغم تحقیقات فراوان، شناسایی و ردیابی چشم کماکان به دلایلی همچون انسداد چشم توسط پلک‌ها بسته و یا باز بودن چشم‌ها تغییرات در مقیاس و یا حرکت سر به عنوان یک مبحث چالش‌برانگیز مطرح می‌شود. تحقیقات فراوانی در زمینه‌ی شناسایی و ردیابی چشم و البته تخمین جهت نگاه انسان انجام شده است. یکی از دلایلی که این ردیابی و تخمین را با سختی مواجه کرده است مسئله انسداد است. سایر کاربردهای بینایی ماشین از جمله بازشناسی صورت، ردیابی انسان و کاربردهای پزشکی مختلف نیز ممکن است با مسئله انسداد و یا تغییرات شکل مواجه باشند؛ اما پیچیدگی آن‌ها از نظر اندازه و فرکانس تکرار در مقایسه با چشم ناچیز است.^[۴]

تصویر چشم ممکن است توسط توزیع شدت مردمک عنبیه و قرنیه و یا از طریق شکل توصیف شود. عواملی همچون نژاد، زاویه دید، جهت سر، رنگ، بافت شرایط نوری، موقعیت عنبیه در کاسه چشم و حالت چشم (باز/بسته) به‌اندازه زیادی بر روی ظاهر چشم اثرگذار هستند. شکل چشم ممکن است زمانی که از زوایای مختلف دیده می‌شود تغییر زیادی داشته باشد. برای مثال پلک‌ها ممکن است از یک زاویه صاف و از

¹ occlusion

زاویه دیگر خمیده به نظر نشان می‌دهد که پلک‌ها صاف هستند در حالی که خط

زردرنگ محور اصلی بیضی عنبیه را نشان می‌دهد [۴]



شکل (۱-۲) تغییرات شکل چشم از زوایای مختلف

در این فصل به بررسی کارهای انجام شده در حوزه ردیابی چشم و تخمین جهت نگاه انسان می‌پردازیم. ابتدا طبقه‌بندی کلی انواع روش‌های ردیابی چشم را مطرح کرده و به بررسی کارهای انجام شده در هر دسته می‌پردازیم، سپس تحقیقات جدید را که با این پایان‌نامه ارتباط بیشتری دارند مطرح خواهیم کرد.

در طبقه‌بندی انواع روش‌های ردیابی چشم را می‌توان در سه دسته قرارداد: مبتنی بر شکل، مبتنی بر ظاهر و روش‌های ترکیبی. روش‌های مبتنی بر شکل خود به دو دسته شکل ثابت^۱ و شکل قابل تغییر^۲ تقسیم می‌شوند. روش‌های مذکور از روی ویژگی‌های نقطه محلی چشم و صورت و یا از روی کانتورهای اشان ساخته می‌شوند. ویژگی‌های مربوطه شامل لبه‌ها گوشه‌های چشم یا نقاطی است که بر اساس پاسخ فیلترهای به خصوصی انتخاب شده‌اند. لیمبوس و مردمک از جمله ویژگی‌های رایج هستند. روش‌های مبتنی بر ظاهر^۳ بر مدل‌هایی متکی‌اند که مستقیماً بر اساس ظاهر ناحیه چشمی ساخته شده‌اند. رویکرد مبتنی بر ظاهر (رویکرد جامع) به‌طور مفهومی به انتباطک الگو مرتبط می‌شود. این ارتباط از طریق ساخت یک مدل تصویری و اعمال ردیابی چشم به‌وسیله انتباطک الگو و با کمک یک معیار شباخت میسر می‌شود. روش‌های مبتنی بر ظاهر خود به دو دسته روش‌های شدت و زیر فضای^۴ تقسیم می‌شوند. روش‌های مبتنی بر شدت مستقیماً از شدت و یا شدت فیلتر شده‌ی تصویر به عنوان مدل استفاده می‌کنند در حالی که روش‌های زیر فضای فرض می‌کنند اطلاعات مهم تصویر چشم در زیر فضایی با ابعاد کمتر تعریف شده است روش‌های ترکیبی نیز رویکردهای

¹ Fixed shape

² Deformable shape

³ Appearance-based methods

⁴ subspace

ویژگی شکل و ظاهر را ترکیب کرده تا از مزایای آن‌ها بهره برد.

۲-۲-۲- رویکردهای مبتنی بر شکل

زمانی که چشم باز است می‌توان به خوبی آن را از طریق شکلش توصیف کرد شکلی که شامل کانتورهای عنبیه و مردمک و همچنین پلک‌ها می‌شود. تقسیم‌بندی رویکردهای مبتنی بر شکل به این برمی‌گردد که مدل انتخابی بیضوی ساده و یا مدل‌های پیچیده‌تری در نظر گرفته شود. مدل‌های شکل معمولاً دو بخش دارند: یک مدل هندسی از چشم و یک معیار شباهت. پارامترهای مدل هندسی میزان تغییرات قابل قبول در الگو را نشان می‌دهند. مدل‌های شکلی تغییرپذیر معمولاً بر روی یک الگوی تغییرپذیر کلی استوارند که در آن توسط تغییر در مدل شکل با کمک کمینه‌سازی انرژی مکانیابی چشم انجام می‌شود. یک ویژگی مهم این روش‌های این است که توانایی تطابق با تغییرات شکل مقیاس و همچنین چرخش را دارد.

مدل‌های شکل بیضوی ساده

بسیاری از کاربردهای ردیابی چشم تنها به تشخیص و ردیابی عنبیه و یا مردمک نیاز دارند. بر اساس زاویه دید هردوی عنبیه و مردمک بیضوی دیده شده و بنابراین می‌توانند توسط پنج پارامتر شکلی توصیف شوند. مدل‌های بیضوی ساده شامل روش‌های مبتنی بر رأی‌گیری^۱ و روش‌های برازش مدل^۲ می‌شوند. روش‌های رأی‌گیری ویژگی‌هایی را انتخاب می‌کنند که یک فرضیه داده شده را از طریق رأی‌گیری یا فرآیند اجتماع گیری تأیید می‌کنند؛ در حالی که رویکردهای برازش مدل ویژگی‌های انتخاب شده را به مدل برازش^۳ می‌کنند (مانند بیضی). در کارهایی همچون^[۵] از آستانه‌های شدت تصویر با تخمین مرکز بیضی مردمک استفاده شده است. روش‌های تشخیص لبه برای استخراج لیمبوس و یا حاشیه مردمک استفاده می‌شوند. بخش‌های مختلفی در تصویر ممکن است پروفایل شدت مشابهی با نواحی عنبیه و مردمک داشته باشند و - بنابراین فقط در تنظیمات محدودی قابل استفاده‌اند.

تبديل هاف^۴ نیز می‌تواند برای استخراج مناسب عنبیه و یا مردمک به کار رود اما این روش نیازمند

¹ Voting based

² Model fitting

³ fit

⁴ Hough transform

ردیابی ویژگی صریح است. معمولاً به دلیل بهره‌وری یک محدودیت شکل مدور اعمال شده و بنابراین مدل فقط برای حالات صورت از جلو کار می‌کند. توجه به این حقیقت که تغییر عنبیه می‌تواند با دو درجه آزادی (متناظر با Pan و Tilt) مدل شود می‌تواند منجر به کاهش بار محاسباتی شود. در یک روش رأی‌گیری دیگر پیشنهاد شده است که از اطلاعات زمانی و مکانی برای ردیابی مکان چشم استفاده می‌کند. در این مقاله از گرادیان استفاده شده است با دانستن این موضوع که گرادیان در طول محدوده عنبیه از مرکز عنبیه به بیرون است.^[۴]

روش رأی‌گیری مشابهی نیز در [۶] استفاده شده است. این مقاله بر پایه انحنا ایزووفوت^۱ در شدت تصویر بنا شده و مستقیماً از جهت لبه در فرآیند رأی‌گیری استفاده می‌کند. این رویکرد برای کاهش تعداد مثبت‌های اشتباه^۲ بر یک مدل صورت پیشین و میانگین انتروپومورفیک متکی است. به دلیل این‌که این روش‌ها به بیشینه در فضای ویژگی متکی‌اند ممکن است زمانی که تعداد ویژگی‌ها در ناحیه چشم کاهش می‌یابد، ویژگی‌های دیگری را برای چشم به اشتباه به شمار آورند. (مانند ابرو و یا گوش‌های چشم). این روش‌ها معمولاً زمانی استفاده می‌شوند که یک ناحیه جست‌وجو محدود در دسترس است.

نویسنده‌گان [۷] نیز عنبیه را به عنوان یک بیضی مدل کرده‌اند اما این بیضی به‌طور محلی به تصویر برآش شده است. این برآش از طریق بهینه‌سازی‌های EM و RANSAC انجام شده است.

در [۸] ردیابی عنبیه و گوش‌های چشم با استفاده از دستگاه پوشیدنی و یک نور دهنده مادون‌قرمز با برد کم^۳ استفاده شده است.

مدل‌های شکل ساده معمولاً کارا هستند و می‌توانند ویژگی‌هایی از جمله عنبیه و مردمک را تحت زوایای مختلف دید مدل کنند. اگرچه مدل‌های ساده توانایی نمایش تغییرات ویژگی‌هایی از جمله پلک‌ها گوش‌های چشم و ابروها را ندارد. معمولاً تصاویر و آستانه‌هایی با کنتراست بالا برای استخراج ویژگی استفاده می‌شوند.

مدل‌های شکل پیچیده

مدل‌های شکل پیچیده همان‌طور که از نامشان پیداست برای مدل‌سازی با جزئیات بیش‌تر از شکل چشم به کار می‌رond. یک مثال از این مدل‌ها مدل الگو قابل تغییر^۴ ارائه شده در [۹] است. مدل چشم قابل تغییر شامل دو مقطع مخروطی است که نماینده پلک‌ها هستند (توسط یازده پارامتر مدل می‌شوند) و همچنین دایره‌ای که عنبیه را نشان می‌دهد. نتایج تجربی نشان می‌دهد مکان اولیه الگو مهم است. برای مثال زمانی که شروع

¹ iosphere

² False positive

³ Near infrared

⁴ Deformable-template

با الگویی بالای ابروها باشد الگوریتم در شناسایی چشم‌ها شکست می‌خورد.

در [۱۰] نویسنده‌گان از ردیابی مرکز مردمک به کمک روش شکل قابل تغییر و مبتنی بر شدت استفاده کرده‌اند. در این روش از الگوریتم^۱ DAISMI برای تشخیص کاسه چشم و از الگوریتم^۲ DTBGE به عنوان یک فیلتر نویز استفاده شده است. DAISMI از تصاویر مقیاس-خاکستری استفاده کرده و به جستجوی تیره‌ترین کانتور و یا شکل می‌پردازد.

مشکل دیگر پیچیدگی توصیف الگوها است. به علاوه رویکرد مبتنی بر الگو زمانی که به دلیل نزدیکی پلک‌ها یا حالت سر غیرمستقیم (از جلو نباشد) انسداد چشم وجود دارد احتمالاً با مشکل مواجه می‌شود. روش‌های مبتنی بر الگوی تغییرپذیر منطقی و به طور کلی دقیق و عمومی هستند اما از برخی محدودیت‌ها رنج می‌برند: (۱) از نظر محاسباتی سنگین‌اند (۲) ممکن است تصاویر با کنتراست بالا نیاز داشته باشند (۳) معمولاً برای مکان‌یابی موفق نیاز است نزدیک به محل چشم آغاز^۳ شوند. درنتیجه برای حرکات سر زیاد آن‌ها به روش‌های دیگری برای تأمین این آغاز خوب دارند. (۴) ممکن است نتوانند تغییرات حالات صورت و انسداد چشم‌ها را به خوبی مدیریت کنند. درحالی‌که برخی مدل‌های قابل تغییر از جمله مدل‌های ماری^۴ اجازه تغییرات زیاد در شکل را می‌دهند سایر مدل‌های قابل تغییر تغییرات زیاد در شکل چشم‌ها را برنمی‌تابند.

روش‌های مبتنی بر ویژگی

روش‌های مبتنی بر ویژگی را به منظور تعیین مجموعه‌ای از ویژگی‌های متمایز‌کننده خصوصیات چشم انسان بررسی می‌کنند. لیمبوس مردمک (تصاویر تیره‌اروشن از مردمک) و انکاس قرنیه برخی از ویژگی‌های رایج در مکان‌یابی چشم محسوب می‌شوند. در مقایسه با روش‌های جامع هدف روش‌های مبتنی بر ویژگی تعیین ویژگی‌های محلی و البته حاوی اطلاعات مفیدی^۵ از چشم و صورت است که نسبت به تغییرات روشناهی و زاویه دید کم‌تر حساسیت کمتری داشته باشند.

ویژگی‌های محلی توسط شدت:

ناحیه چشمی شامل چندین مرز است که ممکن است تفاوت در سطح خاکستری قابل شناسایی باشند. در [۱۱] از پرسپترون چندلایه به منظور استخراج ویژگی‌های صورت توسط مکان‌یابی چشم‌ها

¹ Deformable angular integral search by minimum intensity

² Deformable template-based 2D gaze estimation

³ initialization

⁴ Snake-models

⁵ informative

استفاده شده است.

در [۱۲] یک رویکرد ترکیبی برای دسته‌بندی چشم با استفاده از یک الگوریتم تکاملی ارائه شده است تا مجموعه‌ای از ویژگی‌های بهینه (میانگین شدت روشنایی لابلسین و آنتروپی) را بیاید. به جای ردیابی ویژگی‌های چشم در [۱۳] پیشنهاد شده است ناحیه بین دو چشم تشخیص داده شود. ناحیه بین دو چشم نواحی تیره در سمت چپ و راست خود (چشم‌ها و ابروها) و نواحی نسبتاً روشنی در بالا (پیشانی) و پایین (بینی) دارد. این نواحی در بین بسیاری از مردم یکسان بوده در بین بسیاری از زوایا قابل رؤیت بوده و اعتقاد بر این است که پایدارتر و برای ردیابی آسان‌تر از چشم‌ها است. آزمایش‌ها نشان داده است این الگوریتم زمانی که موها بخشی از پیشانی را پوشانده و یا فرد موردنظر عینک بافريم مشکی داشته باشد به مشکل بر می‌خورد.

ویژگی محلی از طریق پاسخ‌های فیلتر:

پاسخ‌های فیلتر برخی ویژگی‌ها را بهبود می‌بخشند؛ بنابراین استفاده از یک بانک فیلتر اگر به خوبی تعریف شود از اهمیت ویژگی‌های نامرتب می‌کاهد. مقدار یک پیکسل در تصویر پس از اعمال فیلتر متناسب است با شباهت آن ناحیه از تصویر به فیلتر؛ بنابراین مناطقی از تصویر که خصوصیت‌های ویژه‌ای دارند می‌توانند از طریق اندازه مشابهت استخراج شوند. در [۴] روش‌هایی به منظور ردیابی چشم با استفاده از فیلترهای خطی و غیرخطی ارائه شده است. آزمایش‌ها نشان می‌دهند استفاده از فیلترهای غیرخطی نرخ شناسایی بهتری نسبت به فیلترهای خطی مبتنی بر لبه سنتی از خود نشان می‌دهند. نویسنده‌گان [۱۴] یک تصویر را با یک فیلتر چرخشی پیچش^۱ کرده تا جهات گرادیان را به دست آورند. بیشترین مقدار پیچش یک کاندیدا برای مرکز عنبیه به شمار می‌رود. سپس از مکافله‌های^۲ تقارن و فاصله برای مکان‌یابی هر دو چشم استفاده می‌شود.

شناسایی مودمک:

زمانی که چشم از فاصله نسبتاً نزدیکی مطالعه شود مردمک یک ویژگی رایج و نسبتاً قابل اعتماد برای ردیابی چشم به شمار می‌آید. مردمک و عنبیه احتمالاً از نواحی اطرافشان تیره‌تر بوده و اگر کنتراست به میزان قابل قبولی بالا باشد می‌توان از اعمال آستانه استفاده کرد. در برخی روش‌ها مانند آنچه در [۴] به آن اشاره شده است از یک الگوریتم آستانه تکرارشونده برای پیدا کردن مکان مردمک‌ها استفاده شده است. این مهم از طریق جست‌وجوی دو ناحیه تاریک انجام شده است. این روش در حضور سایر نواحی تاریک مانند

¹ convolve

² heuristics

ابروها و یا سایه با شکست مواجه می‌شود. حتی اعمال آستانه‌های مشابه نیز برای هر دو چشم کار صحیحی نیست مخصوصاً در حالتی که سر زاویه دارد. ردیابی نواحی تاریک احتمالاً در روش‌هایی مناسب‌تر هستند که به جای نور مرئی از نور مادون قرمز استفاده می‌شود.

بیش‌تر روش‌هایی که تاکنون از آن‌ها نامبرده شد توانایی ردیابی چشمان بسته را ندارند. مقاله [۴] روشی را پیشنهاد می‌دهد تا به وسیله آن چشم ردیابی شده و پارامترهای چشم از طریق مدل دو حالته (چشم‌های باز/بسته) به دست آیند. در این روش با استفاده از الگوریتم Lucas-Kanade پلک‌ها و گوشه چشم‌ها استخراج می‌شود. لبه و شدت روشنایی عنبیه نیز برای استخراج اطلاعات شکل مشابه با روش الگوی قابل تغییر در [۹] استفاده می‌شوند. این روش برای شناسایی و ردیابی گوشه‌های چشم و به دست آوردن یک تصویر خوب از لبه‌ها نیازمند تصاویری با کیفیت بالا است.

در [۱۵] نیز از انعکاس قرنیه^۱ برای آزمون جراحان استفاده شده است. بدین ترتیب که از روی انعکاس قرنیه جهت نگاه ۱۵ جراح حرفه‌ای با ۱۰ جراح مبتدی مقایسه شده و در نهایت این نتیجه حاصل شده است که جراحان حرفه‌ای مدت‌زمان کمتری جراحی موردنظر را انجام داده‌اند و یکی از دلایل آن متمرکز شدن بر روی نقطه هدف در جراحی است حال آن که جراحان تازه‌کار مقداری از زمان خود را علاوه بر تمرکز بر روی نقطه هدف بر ردیابی ابزارهای جراحی صرف کرده‌اند.

در [۱۶] از تشخیص جهت نگاه در محیط بازی استفاده شده است. در این مقاله از هیچ نوع اتصالات پوشیدنی و یا منابع نوری اضافی استفاده نشده است. هم‌چنین از پدیده انعکاس نور نیز استفاده نشده است. نویسنده‌گان در این مقاله تنها از یک دوربین در مقابل صورت استفاده کرده‌اند تا بتوانند با استفاده از گوشه‌های چشم و تشخیص مردمک آن در تشخیص جهت نگاه بهره ببرند.

روش‌های مبتنی بر ویژگی معمولاً استحکام^۲ خوبی را در برابر تغییرات روشنایی از خود نشان داده‌اند. این روش‌ها در شرایط محیط‌های داخلی و حتی در محیط‌های تاریک بهتر عمل می‌کنند چرا که با قرار گرفتن در محیط‌های بیرونی و روشن‌تر مردمک چشم کوچک‌تر می‌شود. [۱۷]

روش‌های مبتنی بر ظاهر

روش‌های مبتنی بر ظاهر تحت عنوان الگوی تصویر و یا روش‌های جامع نیز شناخته می‌شوند. این روش‌ها مستقیماً و بر اساس ظاهر فوتومتریک چشم‌ها را شناسایی و ردیابی می‌کنند. این ظاهر فوتومتریک توسط توزیع رنگ و یا پاسخ‌های فرکانسی چشم و نواحی اطراف آن توصیف می‌شوند. این روش‌ها مستقل از

¹ Cornea Reflection

² robustness

شیء موردنظر بوده و در اصل توانایی مدل‌سازی سایر اشیا را در کنار چشم‌ها دارد. رویکردهای مبتنی بر ظاهر در دامنه مکانی و دامنه تبدیل یافته^۱ اعمال می‌شوند. یکی از مزایای ردیابی چشم (و در حالت کلی شناسایی اشیا) در دامنه تبدیل یافته، کم کردن تغییرات مربوط به روشنایی است. این امر از طریق حذف باندهایی از فرکانس که نسبت به تغییرات نور حساس‌اند انجام می‌شود. اگرچه در عمل این روش‌ها نسبت به تغییرات معتل در روشنایی مقاوم هستند.

دامنه شدت روشنایی:

در [۴] به مقاله‌ای اشاره شده است که نویسنده‌گان از تفریق پس‌زمینه^۲ برای شروع کار ردیاب مبتنی بر همبستگی^۳ استفاده کرده‌اند. در مقاله‌ای دیگر از مدلی متشکل از دو ناحیه با شدت روشنایی یکنواخت استفاده شده است. یک ناحیه برای منطقه تاریک عنبیه و ناحیه دیگر برای منطقه سفید صلبیه^۴. در پژوهشی دیگر که در همان مرجع معرفی شده است چشم‌ها را با استفاده از ماشین بردار پشتیبان ردیابی می‌کند. کرنل‌های چندجمله‌ای درجه دوم بهترین عملکرد تعمیم را در این پژوهش نشان داده‌اند.

پاسخ‌های فیلتر:

روش‌هایی که از پاسخ فیلتر برای مدل ظاهر استفاده می‌کنند از مقادیر پاسخ مستقیماً و بدون انتخاب این که کدام ویژگی‌ها را انتخاب کنند استفاده می‌کنند. ویژگی‌های چشمی ایده‌آلی که در [۱۸] استفاده شده است ویژگی‌های هار^۵ هستند. ردیاب صورت مقاله [۱۹] مجموعه ویژگی متمایز‌کننده هار را برای تشخیص صورت توسط آدابوست^۶ یاد می‌گیرد. روش‌های مشابه دیگری نیز در [۴] برای ردیابی چشم استفاده شده است.

مهمنترین مزیت استفاده از ویژگی‌های هار بهره‌وری آن‌ها از نظر محاسبات است. اگرچه محاسبه ویژگی‌های هار آسان است کارایی متمایز‌کننده‌شان ممکن است محدود باشد. در [۵] به پژوهش دیگری اشاره شده است که ویژگی جداکننده غیر پارامتری بازگشتی را برای تشخیص چشم و صورت پیشنهاد کرده است که از آدابوست نیز برای آموزش استفاده می‌کند. این روش بر معایب استفاده از ویژگی‌های هار چیره

¹ Transformed domain

² Background subtraction

³ correlation

⁴ sclera

⁵ Haar

⁶ Adaboost

شده است. نویسنده‌گان این مقالات نتایج ردیابی خوبی را برای مکان‌یابی مردمک و در شرایط استفاده از تعداد کمتری ویژگی‌های تفکیک‌کننده گزارش کرده‌اند.

۳-۲-۲- مدل‌های ترکیبی

هدف روش‌های ترکیبی، ترکیب مزایای روش‌های مختلف در یک سیستم است به طوری که بر معایب این روش‌ها نیز غلبه کند.

شکل و شدت:

ترکیب شکل و ظاهر ممکن است برای مثال در روش‌های مبتنی بر بخش^۱ دیده شود. مدل‌های مبتنی بر بخش سعی می‌کنند یک مدل کلی با استفاده از یک مدل شکل برای مکان یک قسمت خاص از تصویر به دست آورند. در این روش‌ها ناحیه چشم ابتدا از طریق آستانه گیری و جستجوی دودویی مشخص شده و سپس به چند بخش شامل ناحیه چشم صلبیه عنبیه تقسیم می‌شود. محدودیت روش‌های مبتنی بر بخش این است که آن‌ها شدت‌های تصویر را مستقیماً بر روی نواحی غیر تکه^۲ مدل نمی‌کنند.^[۴]

[۲۰] روش‌هایی را پیشنهاد می‌کنند که مدل‌های شکل و ظاهر از طریق مدل ظاهر فعال^۳ (AMM) [۲۱] باهم ترکیب می‌شوند. در این مدل‌ها هم‌شکل و هم ظاهر در یک مدل تولیدی^۴ باهم ترکیب می‌شوند. این مدل سپس می‌تواند با تغییر در پارامترها و با توجه به مدل قابل تغییر فراگرفته شده به تصویر برآذش شود.

رنگ و شکل:

می‌توان گفت توزیع رنگ در ناحیه چشمی با اطراف آن تفاوت دارد. علیرغم وجود این حقیقت توجه اندکی به مدل‌های رنگی از چشم شده است البته مواردی از جمله^[۲۱, ۲۳] وجود دارند که از توزیع رنگ برای مدل‌سازی ناحیه چشم استفاده کرده‌اند. در [۲۱] نویسنده از یک مدل رنگ برای ردیاب رنگ نوبت میانگین^۵ برای ردیابی درشت-مقیاس^۱ و یک مدل ظاهر فعال مقیاس برای مکان‌یابی دقیق استفاده کرده

¹ Part-based

² Non-patch

³ Active Appearance Model

⁴ generative

⁵ Mean shift color tracker

است. یک مدل ظاهر فعال مبتنی بر رنگ نیز آزمایش شده است اما کارایی را بهبود نداده است. محدودیت‌های این روش آن است که دو مدل از هم جدا بوده و مدل ظاهر فعال بر روی نتیجه مستقل از ردیاب رنگ است.

۳-۲- سایر روش‌ها

نویسندهاند [۲۴] روشی را ارائه کرده‌اند که در آن مکان‌یابی مردمک چشم بر اساس گروهی از درخت‌های رگرسیون تصادفی^۲ انجام شده است. در این روش تنها یک دوربین کالیبره نشده نیاز است و این امر موجب شده است به عنوان یک روش کاملاً غیرتهاجمی از آن نامبرده شود. در این روش فرض شده است مکان تقریبی چشم‌ها با استفاده از ردیاب‌های صورت موجود به صورت نواحی مستطیلی که حاوی چشم‌ها هستند در دسترس‌اند. سپس مختصات مردمک در این نواحی مشخص می‌شود.

[۲۵] روشی را ارائه می‌دهد که در آن مدل تخمین جهت نگاه نامتغیر با حالت سر و برای دوربین RGB-D دور^۳ است. این روش از تشخیص ویژگی‌های رویکردهای هندسی که نیازمند تصاویر با کیفیت بالا هستند اجتناب می‌کند. در این روش که تخمین جهت نگاه تولیدی هندسی^۴ نامیده می‌شود از مدل گرافیکی استفاده شده است. در این مدل گرافیکی هر پارامتر هندسی به عنوان یک متغیر تصادفی به شمار می‌آید.

[۲۶] فتحی و همکارانش در دانشگاه استنفورد روشی ابتکارانه را ارائه کرده‌اند که اولاً در ویدئوهای اول شخص^۵ انجام شده و ثانیاً بر اساس فعالیتی که فرد موردنظر انجام می‌دهد و به با ترکیب حرکات سر و مکان دست‌ها جهت نگاه او استخراج می‌شود. در این پژوهش باید فرد از دوربینی استفاده کند که به سر او بسته شده است. مدل گرافیکی استفاده شده در این پژوهش در شکل (۲-۲) آمده است:

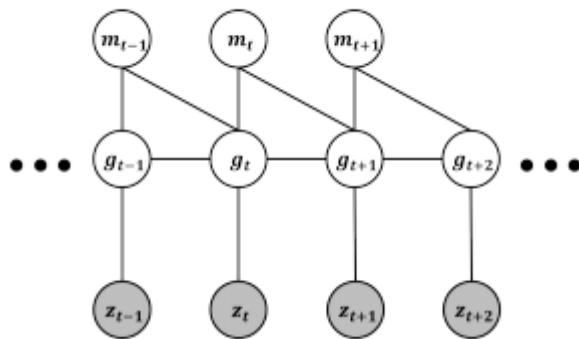
¹ Coarse-scale

² Ensemble of Randomized Regression Trees

³ Distant RGB-D cameras

⁴ Geometric generative gaze estimation

⁵ egocentric



شکل (۲-۲) مدل گرافیکی استفاده شده در [۲۶]

در این مدل g_t جهت نگاه کاربر را در لحظه t بردار ویژگی در لحظه t و m_t صفر و یا یک است؛ که اگر یک باشد بدان معناست که چشم برای نقطه g_t در وضعیت ثابت^۱ قرار دارد. با استفاده از این مدل رابطه زیر برای محاسبه احتمال شرطی جهت نگاه به شرط داشتن بردار ویژگی استخراج شده است:

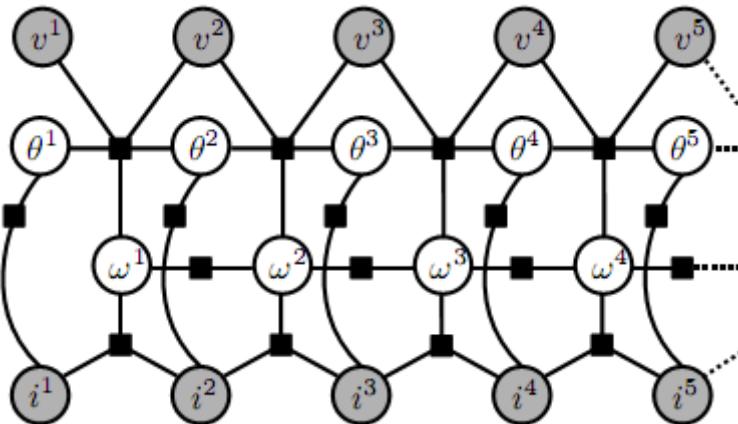
$$P(\{g_t, m_t\}_{t=1}^K | \{z_t\}_{t=1}^K) = \prod_{t=1}^K P(g_t | z_t) \prod_{t=1}^K P(m_t | g_{N(t)}) \quad (1-2)$$

در این رابطه $g_{N(t)}$ طبق مدل گرافیکی همسایگان زمانی فریم t هستند. در [۲۷] ابتدا در یک ویدئو صورت انسان به منظور استخراج نواحی چشم تشخیص داده می‌شود. سپس با ترکیب انرژی شدت و استحکام لبه^۲ مرکز عنبیه بدست آمده و با استفاده از ردیاب گوشه چشم گوشه‌ها استخراج می‌شوند. در این پژوهش همچنین از یک مدل سر سینوسی برای شبیه‌سازی شکل سه بعدی سر استفاده شده است.

در [۲۷] ابتدا در یک ویدئو صورت انسان به منظور استخراج نواحی چشم تشخیص داده می‌شود. سپس با ترکیب انرژی شدت و استحکام لبه^۳ مرکز عنبیه به دست آمده و با استفاده از ردیاب گوشه چشم، گوشه‌ها استخراج می‌شوند. در این پژوهش همچنین از یک مدل سر سینوسی برای شبیه‌سازی شکل سر استفاده شده است.

¹ fixation² Edge strength³ Edge strength

در [۲۸] با استفاده از مدل گرافیکی^۱ CRF رابطه‌ای بین حرکت سر جهت حرکت بدن و ظاهر صورت برقرار شده است تا از آن برای تخمین جهت نگاه استفاده شود. در شکل زیر می‌توان رابطه این پارامترها را در مدل گرافیکی مشاهده نمود.



شکل (۳-۲) گراف فاکتور استفاده شده در [۲۸]

که در آن θ_x^t زاویه سر تخمین زده در لحظه t ، ω^t سرعت زاویه‌ای حرکت سر بین دو لحظه t و $t+1$ اطلاعاتی است که از تصویر لحظه t بدست آمده است و v^t نیز سرعت حرکت فرد را در ویدئو نشان می‌دهد. از رابطه زیر در مدل گرافیکی CRF به عنوان تخمین زننده جهت نگاه استفاده شده است :

$$p(\theta, \omega | i, v) = \frac{1}{Z(x)} \prod_{C_p \in \mathcal{C}} \prod_{\psi_c \in C_p} \psi_c(i_c, v_c, \theta_c, \omega_c) \quad (2-2)$$

که در آن $Z(x)$ تابع بخش‌بندی^۲ بوده که برای نرمال کردن به کار می‌رود.

۴-۲ - کارهای مرتبط

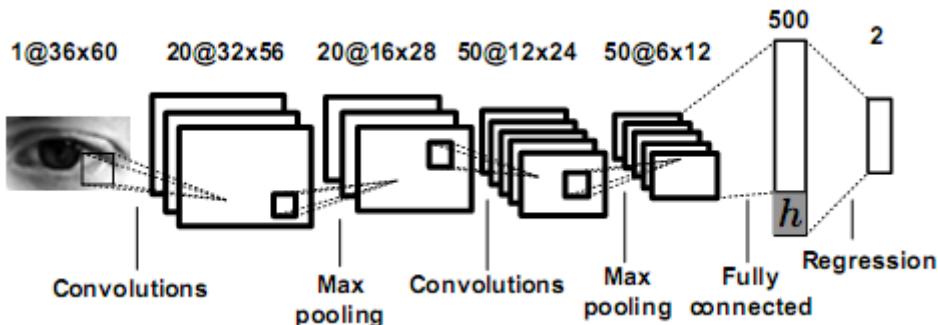
در این بخش به بررسی مقالاتی می‌پردازیم که ارتباط بیشتری با این پایان‌نامه دارند.

¹ Conditional Random Field

² Partition Function

۲-۱-۴-۲- تشخيص جهت نگاه از روی ظاهر در محیط طبیعی

در اين مقاله [۲۹] هدف يادگيري نگاشتي از اطلاعات ورودي به جهت نگاه انسان است. در مقاله مذكور از شبکه LeNet بهمنظور استخراج ويزگي‌هاي عميق از چشم استفاده شده است. معماری شبکه شامل يك لاييه پيچش، يك لاييه ادغام ماکريم، يك لاييه پيچش و يك لاييه ديجير ادغام ماکريم است. شبکه مورداستفاده در اين آزمایش را در شکل (۴-۲) مشاهده می‌کنيد:



شکل (۴-۲) معماری شبکه عميق استفاده شده در [۲۹]

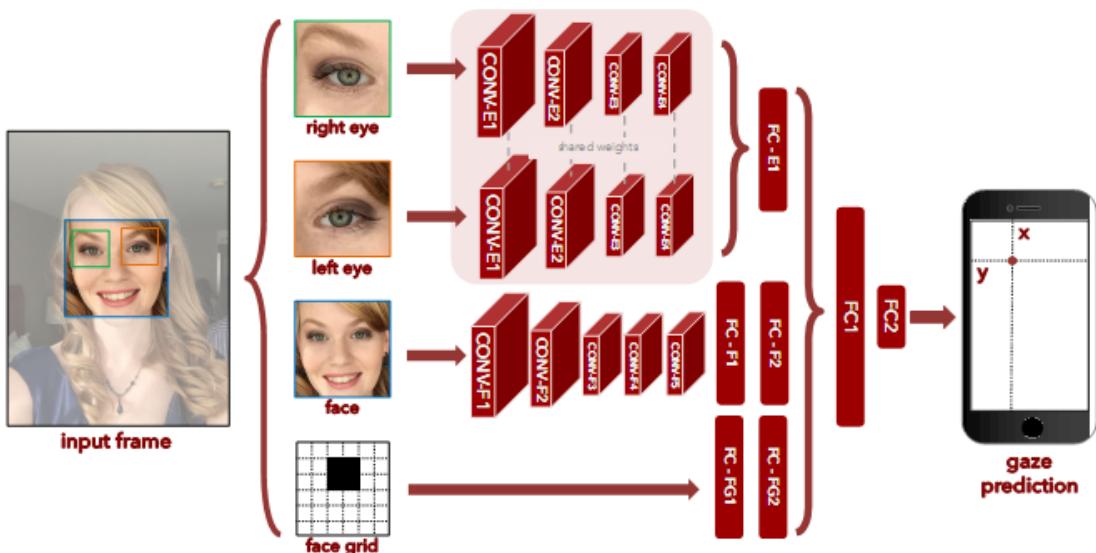
همان‌طور که در شکل (۴-۲) مشاهده می‌شود ورودي اين شبکه يادگيري عميق تصاویر چشم به ابعاد ۳۶ در ۶۰ پикسل است. ورودي ديجير اين شبکه اطلاعات سر (h) است که با خروجی شبکه LeNet ترکيب شده است تا پس از آموزش يك رگرسيون خطی خروجی نهايی را به دست دهد. برای دو لاييه پيچش ابتدائي اندازه فیلتر ۵*۵ پикسل و تعداد اين ويزگي‌هاي به ترتيب ۲۰ و ۵۰ است. تعداد واحدهای نهاي در لاييه تماماً متصل ۵۰۰ عدد است. در اين لاييه تمامی واحدها به تمامی نقشه‌های ويزگي لاييه پيچش قبل متصل هستند. در اين مقاله (تابع زيان اقلیدسي) L2 به عنوان تابع زيان استفاده شده است. اين تابع مقدار اختلاف بين زاويه پيش‌بيين شده و زاويه واقعی را در برچسب مجموعه دادگان محاسبه می‌کند. در اين مقاله از مجموعه دادگان MPIIGAZE استفاده شده است.

۲-۴-۲- رديابي چشم برای همه

به گفته نويسندها [۳۰] هدف اصلی از ارائه اين مقاله فراهم‌سازی امكان رديابي چشم در نرم‌افزاری است که بر روی تلفن‌های همراه و تبلتها، بدون نياز به حسگرهای ديجير قابل استفاده باشد. از اين‌رو آن‌ها

۱۴۷۴ مجموعه دادگانی را با عنوان GazeCapture ارائه کرده‌اند. این مجموعه دادگان با کمک شرکت‌کننده جمع‌آوری شده شامل ۲,۴۴۵,۵۰۴ فریم است. از این میان ۲,۱ میلیون فریم در دستگاه‌های آیفون و ۳۶۰ هزار عدد دیگر با استفاده از آی‌پد ضبط شده‌اند.

آن‌ها با استفاده از این مجموعه دادگان یک شبکه عصبی با نام iTracker را آموزش داده‌اند. آن‌ها عملکرد شبکه خود را با توان پردازشی ۱۰-۱۵ فریم در ثانیه و با خطای ۲,۵۳ و ۱,۷۱ سانتی‌متر، به ترتیب برای گوشی تلفن‌همراه و تبلت گزارش کرده‌اند. شبکه مورداستفاده در [۳۰] را در شکل (۵-۲) مشاهده می‌کنید.



شکل (۵-۲) معماری شبکه عمیق استفاده شده در [۳۰]

ورودی این شبکه چشم‌ها، صورت و جای صورت در فریم اصلی تحت عنوان شبکه صورت^۱ است. علیرغم اینکه در تصویر صورت چشم‌ها نیز حضور دارند، چشم‌ها به‌طور جداگانه به این شبکه داده شده‌اند تا با در اختیار داشتن تصاویری با وضوح بالاتری از چشم‌ها، امکان تشخیص تغییرات اندک نیز فراهم باشد.

تصاویر ورودی، اعم از صورت و چشم‌ها، در ابعاد ۲۲۴*۲۲۴ پیکسل به شبکه داده می‌شوند. شبکه صورت یک ماسک دودویی است که مکان و اندازه سر را در فریم ورودی مشخص می‌کند. خروجی شبکه ارائه شده، فاصله به سانتی‌متر از مبدأ دوربین است. اندازه فیلتر برای هر کدام از لایه‌ها به شرح زیر است: CONV-E1,CONV-F1: 11 × 11, CONV-E2,CONV-F2: 5 × 5, CONV-E3,CONV-F3: 3 × 3, CONV-E4,CONV-F4: 1 × 1,

¹ face grid

تعداد نقشه ویژگی برای فیلترهای استفاده شده نیز برای لایه‌های پیچش به شرح زیر است:

CONV-E1,CONV-F1: 96, CONV-E2,CONV-F2: 256, CONV-E3,CONV-F3: 384, CONV-E4,CONV-F4: 64

اندازه لایه‌های تماماً متصل نیز در این شبکه عبارت‌اند از:

FC-E1: 128, FC-F1: 128, FC-F2: 64, FC-FG1: 256, FC-FG2: 128, FC1: 128, FC2: 2

آن‌ها همچنین تأثیر استفاده از چشم‌ها، تصویر صورت و شبکه صورت را به‌طور جداگانه بررسی کرده‌اند.

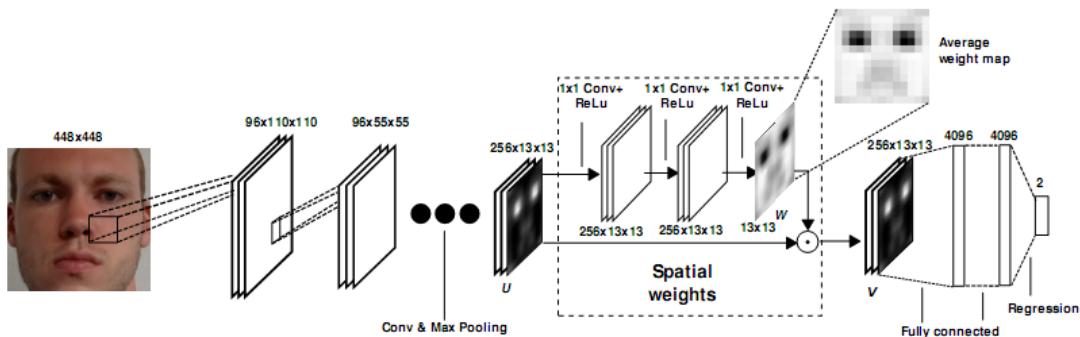
جدول (۱-۲) جدول مقایسه‌ای تأثیر حذف زیر بخش‌های معماری پیشنهادی در [۳۰]

Model	Mobile phone error	Tablet error
iTracker (no eyes)	2.11	3.40
iTracker (no face)	2.15	3.45
iTracker (no fg.)	2.23	3.90

همان‌طور که در جدول (۱-۲) مشاهده می‌کنید تأثیر حذف شبکه صورت بیشتر از سایر عوامل حتی از چشم‌ها بوده است به‌طوری‌که با حذف چشم‌ها از شبکه خطای ۲,۱۱ در حالت تلفن‌همراه و ۳,۴۰ در حالت تبلت بوده است، حال آن‌که حذف کردن شبکه صورت خطای ۲,۲۳ و ۳,۹۰ را به ترتیب در تلفن‌همراه و تبلت ایجاد کرده است.

۳-۴-۲- تشخیص جهت نگاه از روی تصویر صورت

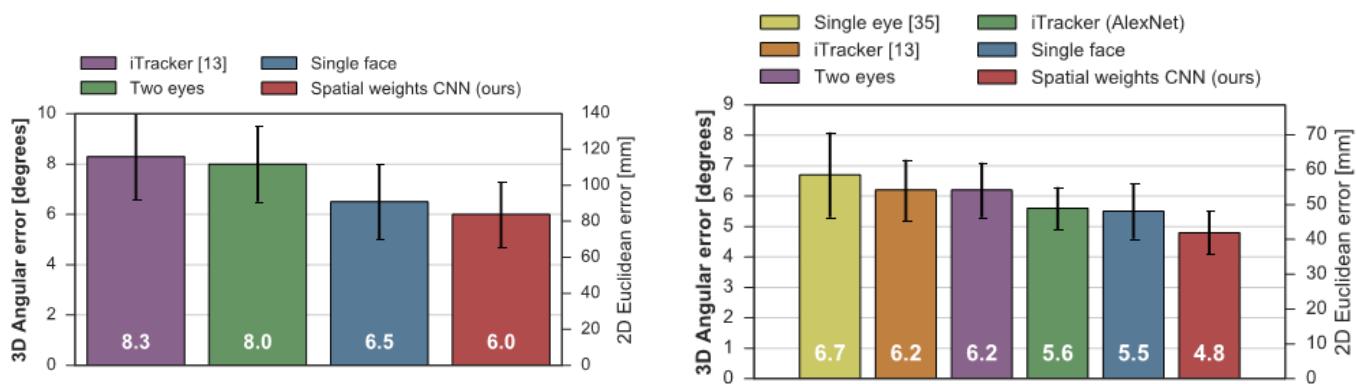
نویسنده‌گان این مقاله [۳۱] علاوه بر بهره‌گیری از شبکه‌های عصبی پیچشی بر روی تصویر صورت از وزن‌های مکانی استفاده کرده‌اند که بر روی نقشه‌های ویژگی اعمال شده‌اند تا اطلاعات بخش‌های مختلف صورت به‌خوبی در یادگیری لحاظ شوند. آن‌ها معتقدند علیرغم این‌که از نظر تئوری تفاوت مشارکت بخش‌های مختلف صورت در تخمین جهت نگاه می‌تواند توسط یک شبکه پیچشی عادی آموخته شود، ارائه روشی که شبکه را مجبور کند این تفاوت را بهتر بیاموزد حائز اهمیت است. در راستای اعمال این نظارت قوی‌تر، آن‌ها از لایه‌های پیچش ۱*۱ پیکسلی استفاده کرده‌اند. شکل (۲-۶) روند این کار را نشان می‌دهد:



شکل (۶-۲) معماری شبکه عمیق پیشنهادی در [۳۱]

همان‌گونه که در شکل (۶-۲) مشاهده می‌شود، ورودی این شبکه تصاویر 448×448 پیکسل صورت است. شبکه پایه استفاده شده همان شبکه Alexnet با ۵ لایه پیچش است با این تفاوت که اندازه نقشه‌های ویژگی تغییر یافته‌اند؛ برای مثال آخرین لایه پیچش پیش از اعمال وزن‌های مکانی $256 \times 13 \times 13$ عدد نقشه ویژگی را به دست می‌دهد. بخش اعمال وزن‌های مکانی از سه لایه پیچش اضافی هرکدام با اندازه فیلتر 1×1 به علاوه یک لایه ReLU تشکیل شده است. اگر ورودی این بخش یک بردار $N \times H \times W$ به دست آمده از شبکه Alexnet قبلی باشد، قسمت بالایی در بخش وزن مکانی یک ماتریس W را با نام W به عنوان خروجی ایجاد می‌کند. این ماتریس W در هر کدام از نقشه‌های ویژگی با اندازه $H \times W$ ضرب می‌شود. بدین ترتیب خروجی بخش وزن مکانی، V ، به اندازه $N \times H \times W$ (در اینجا $256 \times 13 \times 13$) است. در نهایت دو لایه تماماً متصل خروجی موردنظر را به کمک رگرسیون خطی به دست می‌دهند.

شکل (۷-۲) دقیق این روش را در مقایسه با سایر روش‌های ارائه شده نشان می‌دهد:



شکل (۷-۲) مقایسه عملکرد معماری پیشنهادی [۳۱]

در این شکل، کمترین میزان خطأ بر روی مجموعه دادگان MPIIGaze با اندازه ۴,۸ درجه متعلق به

روش ارائه شده است. کمترین خطا بر روی مجموعه دادگان EYEDIAP نیز با ۶ درجه متعلق به همین روش است.

۲-۵- نتیجه‌گیری

در این فصل ابتدا کلیه روش‌های مطرح شده در حوزه ردیابی چشم و تشخیص جهت نگاه با دید کلی بررسی و دسته‌بندی شدند. در انتها تلاش شد برترین کارهای انجام شده بر روی مجموعه دادگان مورداستفاده در این پایان‌نامه که در سال‌های ۲۰۱۵ تا اواخر ۲۰۱۶ ارائه شده بودند موردبحث و بررسی قرار گیرند. در این فصل تلاش شد تا ضمن بررسی کارهای گذشته در این حوزه و دسته‌بندی روش‌های مختلف، جدیدترین راهکارهای مورداستفاده موردنظر بررسی قرار گیرند.

فصل ۳:

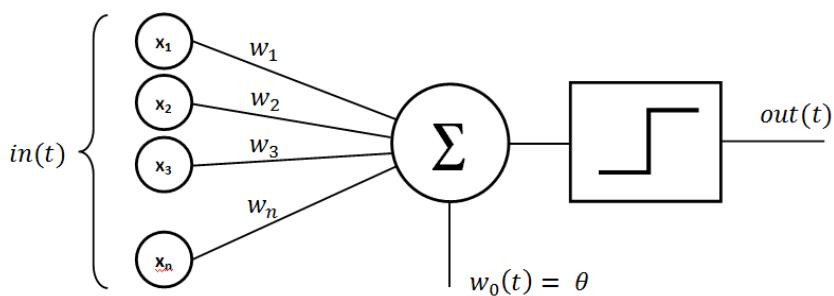
مبانی تحقیق

۱-۳ - مقدمه

در این پژوهش، تمرکز بر حل مسئله تخمین جهت نگاه انسان با استفاده از شبکه‌های عصبی پیچشی و میدان‌های تصادفی شرطی پنهان است. شبکه‌های عصبی پیچشی در حوزه یادگیری عمیق و میدان تصادفی شرطی پنهان در حوزه مدل‌های گرافیکی احتمالاتی دسته‌بندی می‌شوند. بدین منظور در این فصل تلاش شده است ابتدا به معرفی اجمالی این مفاهیم پرداخته شده و سپس پژوهش انجام شده توضیح داده شود.

۲-۳ - شبکه‌های عصبی

بیش از چندین دهه از زمان ساخت اولین ساختار شبکه عصبی مصنوعی مشابه با عملکرد شبکه عصبی انسان می‌گذرد و در طول زمان، ساختارهای ارائه شده توسط محققان، تکمیل شده و سیر تکامل چشمگیری داشته است. در ابتدا، شبکه‌های عصبی از ساختاری تک پرسپترونی تشکیل شده بود [۳۲] (شکل ۱-۳). به مرور زمان، ارتباطات بین پرسپترون‌ها تعریف شد و با توجه به اینکه توانایی پردازشی کامپیوترها نیز افزایش پیدا کرده بود، تحقیق بر روی ارائه معماری‌های مختلف برای شبکه‌های عصبی‌ای که برای مسائل مختلف کاربرد داشته باشند رونق فراوان گرفت.



شکل (۱-۳) نمونه‌ای از ساختار یک پرسپترون ساده [۳۲]

امروزه معماری‌های مختلفی از شبکه‌های عصبی مصنوعی ارائه شده است که با توجه به نحوه تعریف

ارتباطات بین گره^۱های هر لایه، تعداد لایه‌های مخفی موجود، نحوه ارتباطات بین لایه‌ها، توابع فعال‌سازی و غیره قابل جداسازی هستند. با توجه به اینکه شبکه‌های عصبی عمیق، رویکردی جدید و انقلابی را با خود به همراه داشته‌اند، در اینجا شبکه‌های عصبی به دو دسته سنتی و عمیق تقسیم شده و موردنرسی قرار گرفته است.

۲-۲-۳- شبکه‌های عصبی سنتی

استفاده از شبکه‌های عصبی سنتی، یکی از پرکاربردترین گزینه‌ها برای مدل‌سازی و آموزش در سامانه‌های یادگیری ماشین است. در حوزه ردیابی چشم و تخمین جهت نگاه نیز پس از تعریف ویژگی‌های موردنظر طراحان سامانه، با کمک شبکه‌های عصبی و با توجه به این نکته که با افزایش لایه‌های مخفی آن می‌توان تقریباً هر فضای ویژگی پیچیده‌ای را مدل‌سازی کرد، دسته‌بندی‌های مختلفی برای این سامانه‌ها ارائه شده است که برتری‌هایی نیز با توجه به معما ری لایه‌ای خود و توابع فعال‌سازی که در آن‌ها استفاده شده است نسبت به یکدیگر داشته‌اند [۳۳].

۲-۳-۳- شبکه‌های عصبی عمیق

شبکه‌های عصبی عمیق^۲ حوزه جدیدی در شبکه‌های عصبی هستند که اخیراً پیشرفت‌های زیادی را در موضوعات مختلف مرتبط با یادگیری ماشین با خود به همراه آورده‌اند. مشکلات متعددی در عمیق کردن بیش از حد شبکه‌های عصبی سنتی وجود داشته است که از جمله آن‌ها می‌توان به مشکل کاهش گرادیان در لایه‌های عمیق و مشکل بازنثر خطای شبکه عصبی در لایه‌های متعدد معما ری سنتی اشاره کرد. با توجه به ظهور حوزه جدیدی از یادگیری ماشین به نام یادگیری عمیق که با خود تعریف جدیدی را برای ساختارهای شبکه‌های عصبی عمیق ارائه کرده است، مشکلات مذکور رفع شده و توانایی استفاده از شبکه‌های عمیق با قابلیت‌های مختلف برای محققان فراهم شده است.

ساختارهای مختلفی برای این شبکه‌های عصبی عمیق مطرح شده است که از جمله پرکاربردترین آن‌ها

¹ Node

² Deep Neural Network (DNN)

می‌توان به شبکه‌های پیچشی، شبکه‌های بازگشتی^۱ و شبکه‌های باور عمیق^۲ اشاره کرد. هرکدام از ساختارهای اشاره شده، در برخی از مسائل مطرح در حوزه یادگیری ماشین، بیشتر از سایر روش‌ها پیشرفت داشته و مورد توجه قرار گرفته است.

یادگیری عمیق

یادگیری عمیق^۳ شاخه‌ای از بحث یادگیری ماشین و مجموعه‌ای از الگوریتم‌هایی است که تلاش می‌کنند مفاهیم انتزاعی سطح بالا را با استفاده از یادگیری در سطوح و لایه‌های مختلف مدل کنند. یادگیری عمیق در واقع نگرشی جدید به ایده شبکه‌های عصبی است که سالیان زیادی است وجود داشته و هر چند سال یکبار در قالبی جدید خود را نشان می‌دهد.

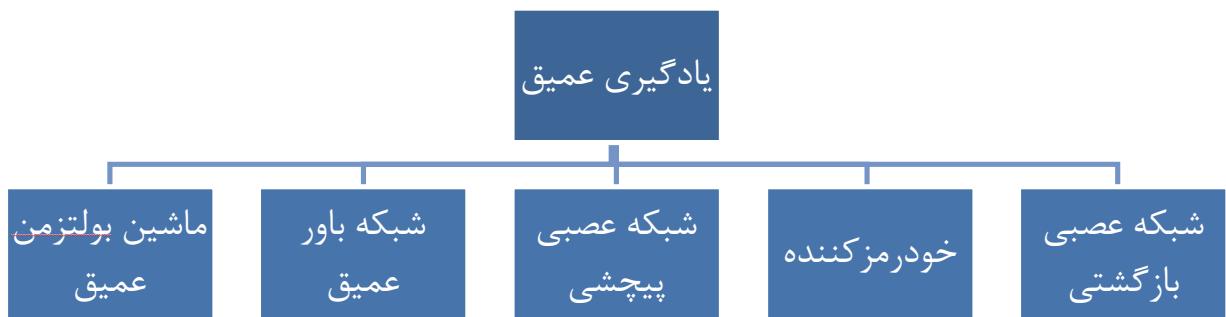
در سال‌های اخیر با افزایش قدرت محاسباتی رایانه‌ها و امکان ایجاد و آموزش این شبکه‌های عمیق با پارامترهای بسیار زیاد، یادگیری عمیق عملکرد مطلوبی را در حوزه‌های مختلف هوش مصنوعی و یادگیری ماشین از خود نشان داده است.

می‌توان سال ۲۰۰۶ را ورود یادگیری عمیق به دایره واژگان مطرح در هوش مصنوعی و یادگیری ماشین دانست. هینتون و همکارانش [34] در سال ۲۰۰۶ با انتشار مقاله خود، الگوریتمی سریع را برای آموزش شبکه‌های باور عمیق معرفی و عملکرد شبکه خود را بر روی دادگان اعداد دستنویس ارزیابی کردند. پس از چند سال و انجام تحقیقات بیشتر در این زمینه، بنجیو در سال ۲۰۰۹ با انتشار مقاله خود با نام «معماری‌های عمیق برای هوش مصنوعی» [۳۵] به تشریح یادگیری عمیق پرداخته و همچنین معماری‌های مطرح را بررسی و نتایج بدست آمده از آن‌ها را گزارش کرد. در سال‌های بعد از انتشار این مقاله، یادگیری عمیق بسیاری از حوزه‌های تحقیقاتی مطرح را در هوش مصنوعی و یادگیری ماشین تحت تأثیر خود قرار داده است. یادگیری عمیق را می‌توان به چند زیرمجموعه کلی همانند آنچه در شکل (۲-۳) آمده است تقسیم کرد [۳۶، ۳۵].

¹ Recurrent Neural Network

² Deep Belief Network

³ Deep Learning



شکل (۲-۳) نموداری کلی از حوزه‌های یادگیری عمیق

بازنمایی یادگیری

یکی از قابلیت‌های مهم شبکه‌های عصبی عمیق یادگیری بازنمایش^۱ است. با استفاده از یادگیری بازنمایش ویژگی‌های موردنیاز برای ماشین به صورت خودکار به دست می‌آیند. بسیاری از روش‌های مرسوم در یادگیری ماشین، مبتنی بر استخراج دستی ویژگی‌ها^۲ هستند. اکثر این روش‌های اعتقاد دارند نمی‌توان از داده‌های طبیعی مانند عکسی، صوت و متن به صورت خام استفاده کرد. برای مثال در پردازش تصویر عملگرهای مختلفی همچون^۳ SIFT و^۴ HOG بر روی عکس اعمال می‌شوند و یا برخی ویژگی‌های فرکانسی استخراج می‌شود. در پردازش گفتار از ضرایب مختلفی مانند ضرایب^۵ LPC یا ضرایب^۶ MFCC استفاده می‌شود. هم‌اکنون بسیاری از تکنیک‌های یادگیری ماشین با استفاده از همین ویژگی‌ها کار می‌کنند. عملکرد الگوریتم مورداستفاده برای یادگیری داده‌ها، وابستگی زیادی به کیفیت و دقت این ویژگی‌های به‌دست‌آمده دارد. انتخاب و استخراج دقیق ویژگی‌ها نیازمند دانش و خبرگی لازم در زمینه مورد مطالعه است.

به فرآیند تبدیل داده‌های خام به ویژگی‌ها قابل استفاده برای الگوریتم یادگیری ماشین استخراج ویژگی‌ها^۷ می‌گویند. با اجرای این فرآیند، بازنمایش داده برای استفاده در الگوریتم یادگیری به‌دست می‌آید. بازنمایش به‌دست‌آمده معمولاً به صورت ضرایب یا بردار ویژگی است. پس از به‌دست آمدن این بردارها، از یک

¹ Representation learning

² Hand craft features

³ Scale-Invariant Feature Transform

⁴ Histogram Of Gradients

⁵ Linear Predictive Coding

⁶ Mel Frequency Cepstral Coefficients

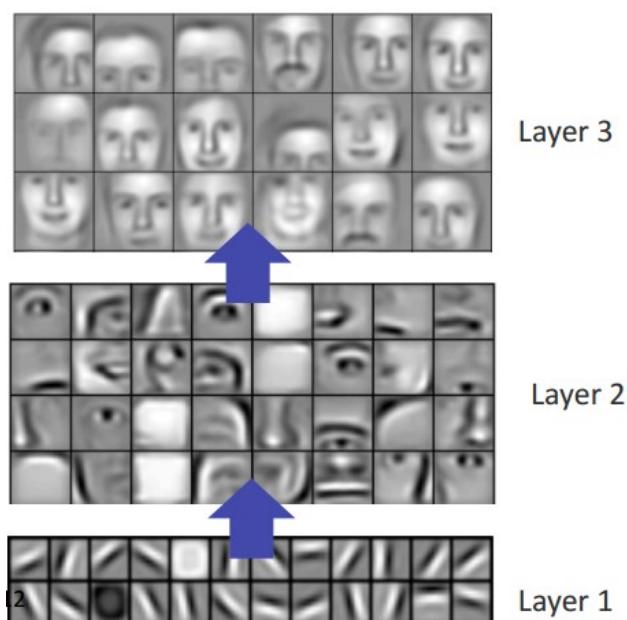
⁷ Feature extraction

یا چند رده‌بند^۱ برای یادگیری استفاده می‌شود. بدین ترتیب نیاز اصلی هر الگوریتم یادگیری، ویژگی‌هایی است که از ورودی‌ها استخراج می‌شود.

یادگیری بازنمایش مجموعه روش‌هایی در یادگیری ماشین است که در آن از داده‌های خام به عنوان ورودی الگوریتم استفاده می‌شود. در این روش‌های بازنمایش مناسب داده‌ها به صورت خودکار به دست می‌آید. با استفاده از تعداد لایه‌های مناسب در یادگیری عمیق می‌توان هر ساختار پیچیده‌ای را به مدل آموزش داد.

یادگیری چندین لایه بازنمایی‌ها

یادگیری عمیق این امکان را به وجود می‌آورد که بتوان مفاهیم با سطح انتزاع بالا را با استفاده از یادگیری چندلایه، از پایین به بالا ساخت شکل ۳-۳.



شکل (۳-۳) بازنمایی ویژگی‌های استخراج شده در لایه‌های مختلف شبکه‌های عمیق [۳۷]

یک مثال رده‌بندی تصاویر را در نظر بگیرید. در رده‌بندی تصاویر، داده‌های ورودی میزان روشناهی پیکسل‌های تصویر است. لایه اول از مدل، وجود لبه‌ها در جهت‌ها و مکان‌های مختلف تصویر ورودی را نشان می‌دهد. لایه بعدی نقش و نگارها و ساختارهای پیچیده‌تر را یاد می‌گیرد. لایه سوم با ترکیب اجزا فراگیری شده در لایه قبلی اجزا اصلی موجود در تصویر ورودی را یاد می‌گیرد. لایه‌های بعدی نیز اشیای موجود در تصویر ورودی را با استفاده از خروجی مرحله قبلی یاد می‌گیرند. نکته اصلی در این نوع یادگیری این است

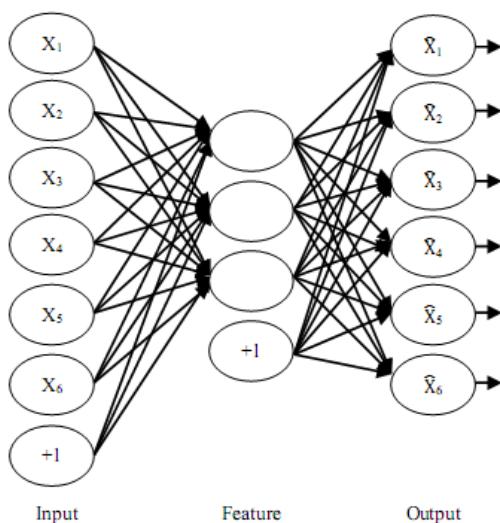
^۱ Classifier

که آموزش ویژگی‌ها به مدل توسط انسان صورت نگرفته بلکه کل فرآیند به صورت خودکار و توسط مدل یادگیری عمیق انجام می‌شود.

در پاسخ به این سؤال که چرا جامعه‌ی علمی اکنون دوباره به این ایده بازگشته است و علت رستاخیز مجدد یادگیری عمیق چیست، می‌توان دلایل متعددی را از جمله به وجود آمدن روش‌های استخراج خودکار مانند خودرمزنده‌ها و همچنین رسیدن به درک بهتری از روش‌های منظم کردن مدل‌ها ذکر کرد. همان‌طور که پیش‌تر بیان شد، یادگیری عمیق دارای زیرمجموعه‌های مختلفی است که معروف‌ترین آن‌ها در نمودار بالا آورده شده‌اند. در ادامه به بررسی زیرمجموعه‌های خودرمزنده، شبکه‌های بازگشتی و شبکه‌های پیچشی پرداخته شده است.

خودرمزنده^۱

خودرمزنده یکی از ساده‌ترین ابزارها برای یادگیری ویژگی‌های همراه با قابلیت جداگاندگی بالا است. یک خودرمزنده سعی می‌کند یک بازنمایی جامع از نمونه‌هایی را که می‌بیند یاد بگیرد. به عبارتی دیگر، خودرمزنده تلاش می‌کند تا یکتابع همانی را یاد بگیرد به‌گونه‌ای که هر ورودی شبکه عصبی را به خود آن ورودی نگاشت کند. بعدازاینکه این شبکه به صورت قابل قبولی آموزش داده شد، خروجی لایه مخفی به عنوان بازنمایی داده ورودی در نظر گرفته می‌شود، به عبارتی دیگر، هنگام استفاده از خودرمزنده، لایه آخر حذف شده و هر داده ورودی به خروجی لایه‌های نهان نگاشت می‌شود. ساختار معماری خودرمزنده در تصویر (۴-۳) قابل مشاهده است.



شکل (۴-۳) ساختار معماری خودرمزنده با یک لایه مخفی [۳۸]

^۱ Autoencoder

تابع هدف تعریف شده برای خود رمزکننده و تعاریف مربوط به آن، به شکل زیر است:

$$J_{Sparse}(W, b) = J(W, b) + \beta \sum_{j=1}^{s_2} KL(\rho || \hat{\rho}_j) \quad (1-3)$$

$$KL(\rho || \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j} \quad (2-3)$$

$$J(W, b) = \left[\frac{1}{m} \sum_{i=1}^m J(W, b; x^{(i)}, y^{(i)}) \right] + \frac{\lambda}{2} \sum_{l=1}^{n_l-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (W_{ji}^{(l)})^2 \quad (3-3)$$

جدول (۱-۳) نمادهای به کاررفته در فرمول‌های (۱-۳)، (۲-۳) و (۳-۳) را معرفی می‌کند.

جدول (۱-۳) نمادهای به کاررفته در خود رمزکننده

نماد	تعریف
x	ویژگی‌های ورودی برای آموزش یک نمونه
y	مقادیر خروجی
$(x^{(i)}, y^{(i)})$	نامین نمونه آموزشی
W	پارامتر مربوط به ارتباط بین واحدهای لایه‌ها
b	مقدار بایاس تعیین شده برای ارتباط بین دو لایه
ρ	پارامتر تنک کردن ^۷ (مشخص کننده سطح موردنظر تنک بودن
$\hat{\rho}_i$	متوسط فعال‌سازی واحد مخفی i در خود رمزکننده
β	وزن تعریف شده برای جریمه تنک شدن سامانه
λ	پارامتر تنظیم کننده

نکته مهم، انتخاب تعداد ویژگی‌ها (تعداد نورون‌های لایه مخفی) است. در صورتی که تعداد نورون‌های لایه مخفی را کم انتخاب کنیم (یعنی کمتر از ابعاد داده ورودی) منجر به کاهش ابعاد داده ورودی می‌شود و رفتاری مشابه با PCA انجام می‌دهد، اما در صورتی که تعداد لایه مخفی را بسیار بزرگ‌تر از ورودی (یا به صورت کلی بزرگ) در نظر بگیریم و محدودیت‌هایی را روی نورون‌های لایه مخفی اعمال کنیم، به

ساختاری بسیار جذاب برای داده‌های ورودی دست پیدا می‌کنیم. همواره محققان به دنبال بازنمایی تنک از داده‌ها بوده‌اند. بازنمایی تنک منجر به عمومیت بیشتر ویژگی‌ها و قابلیت تمایز بالای آن‌ها می‌شوند. در اینجا نیز با انتخاب تعداد بالاتری برای تعداد نورون‌های لایه مخفی و اضافه کردن عبارتی به بهینه کردن گرادیان شبکه به‌گونه‌ای که لایه مخفی را مجبور کند همواره درصد کمی از آن‌ها فعال باشند، می‌توان به بازنمایی تنک داده‌ها به صورت خودکار دست پیدا کرد.

شبکه‌های بازگشتی عمیق^۱

برای مسائلی مانند پردازش گفتار و پردازش زبان طبیعی که داده‌ها به صورت ترتیبی هستند؛ معمولاً استفاده از شبکه‌های عصبی بازگشتی بهتر است [۳۹]. شبکه‌های بازگشتی در هر واحد زمان یک ورودی از دنباله‌ی دادگان را پردازش می‌کنند. سابقه^۲ ورودی‌های قبلی به صورت ضمنی در شبکه ذخیره می‌شود. در شبکه‌های بازگشتی به دلیل وجود بازخورد و داشتن وزن‌های بازگشتی در هر نورون، مفهومی به نام واحد زمان^۳ وجود دارد. استفاده از زمان به صورت گسسته در شبکه‌های عصبی بازگشتی سبب کاهش پیچیدگی در توصیف و آموزش شبکه می‌شود.

شبکه‌های پیچشی^۴

یک شبکه عصبی پیچشی معمولاً از یک یا چند لایه پیچش^۵ تشکیل شده (که گاهی یک لایه ادغام^۶ بعد از آن می‌آید) و سپس چند لایه تمام متصل مانند یک شبکه عصبی چند لایه در انتهای وجود دارد. در شبکه‌های پیچشی عمیق، توابع فعال‌ساز مختلفی از جمله سیگموئید، Tanh و ReLU استفاده می‌شود. ساختار شبکه عصبی پیچشی نسبت به انتقال تغییرناپذیر است. یکی دیگر از ویژگی‌های شبکه عصبی پیچشی، سادگی آن برای یادگیری است زیرا تعداد بسیار کمتری وزن نسبت به یک شبکه عصبی چند لایه تماماً متصل^۷ برای یادگیری وجود دارد.

معمولًا در شبکه‌های عصبی از لایه‌های پیچش و ادغام پشت سر هم استفاده می‌شود. لایه ورودی یک

¹Recurrent Neural Networks

² history

³ Time step

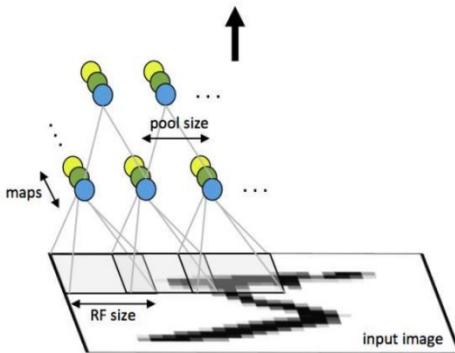
⁴ Convolutional Networks

⁵ convolution

⁶ Pooling

⁷ Fully Connected

تصویر با ابعاد m^*m^*r است که m طول و عرض تصویر و r نیز کانال‌های رنگی^۱ یا نقشه ویژگی‌های تصویر نام دارد. لایه پیچش k فیلتر دارد که اندازه هر فیلتر n^*n^*q است که n کوچک‌تر از ابعاد تصویر است و q نیز می‌تواند به اندازه r یا کمتر باشد. اعمال هر فیلتر یک نقشه ویژگی می‌سازد و می‌توان بعد از آن یک لایه ادغام با اندازه p^*p اعمال کرد. شکل (۳-۵) یک قسمت از یک شبکه عصبی پیچشی را نشان می‌دهد که در بالای تصویر ورودی، لایه پیچش و بر روی آن، لایه ادغام قرار دارد.



شکل (۳-۵) نمونه‌ای از لایه پیچش و ادغام بر روی داده ورودی [۴۰]

در ادامه به بررسی لایه‌های پیچش و ادغام شبکه‌های پیچشی پرداخته شده است. همچنین در انتهای به بررسی توابع فعال‌ساز مناسب شبکه‌های عمیق می‌پردازیم.

پیچش

تصاویر طبیعی دارای ویژگی ایستایی^۲ هستند و به معنای این است که آماره‌های یک قسمت از تصویر با قسمت‌های دیگر تصویر یکسان است. این مطلب بیان‌گر این است که ویژگی‌های به دست آمده از یک قسمت از تصویر می‌تواند در قسمت‌های دیگر تصویر نیز به کار رود و بنابراین می‌توان آن ویژگی‌ها را در همه‌جا به کار برد.

به‌طور دقیق، در صورتی که ویژگی‌هایی از تکه‌هایی که به‌طور تصادفی از تصاویر انتخاب شده است (برای مثال در ابعاد 8^*8)، یاد گرفته شود، آنگاه می‌توان این استخراج کننده ویژگی را بر روی هر قسمت از تصویر اعمال کرد. اگر که هر کدام از این استخراج‌کننده‌های ویژگی را که از یک محدوده 8^*8 توانایی استخراج ویژگی دارد، بر روی یک تصویر بزرگ‌تر به‌طور کامل پیچشی^۳ (درهم‌آمیخته) کنیم آنگاه مقادیر جدیدی

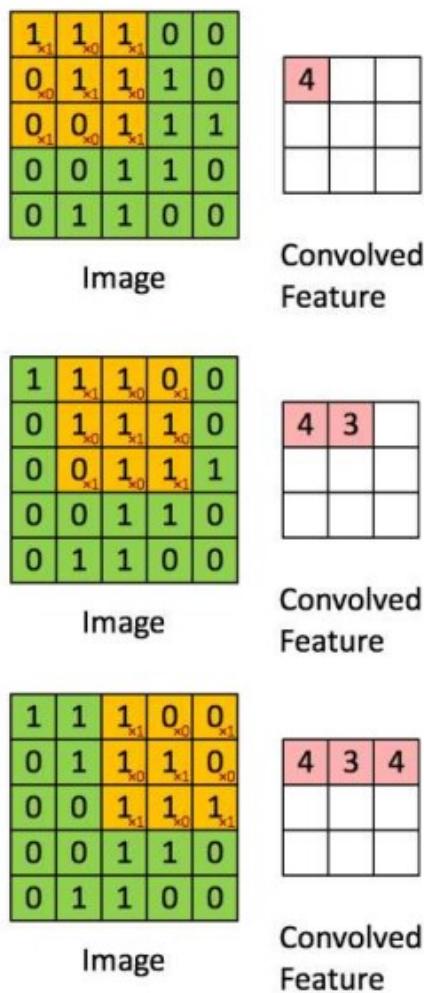
¹ Feature map

² Stationary

³ Convolve

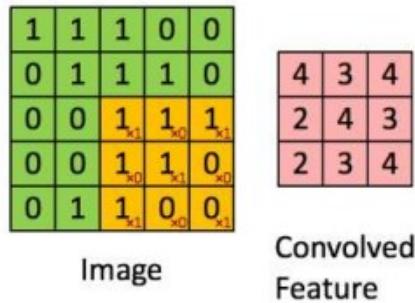
برای تمام نقاط تصویر به دست خواهد آمد.

در شکل (۶-۳) عملیات پیچش در ابعاد بسیار پایین به صورت مرحله به مرحله نشان داده شده است. در اینجا فرض شده است که تصویر کلی با ابعاد 5×5 است و از هر بخش 3×3 از آن، یک ویژگی استخراج می‌شود. این ویژگی با وزن‌های نشان داده شده محاسبه شده و در جدول سمت راست جایگذاری شده است. پس از کامل شدن جدول، تعداد ۹ ویژگی از تصویر به صورت پیچشی شده استخراج می‌شود.



شکل (۶-۳) نمونه‌ای از مراحل عملیات پیچش بر روی داده ورودی [۴۱]

درنهایت پس از آن که ویژگی‌های تمام بخش‌های 3×3 در تصویر استخراج شد، خروجی نمایش داده شده در شکل (۷-۳) حاصل می‌شود که در سمت راست ویژگی‌های پیچشی نهایی نشان داده شده است.



شکل (۳-۷) خروجی نهایی نمونه ورودی پیچشی در [۴۱]

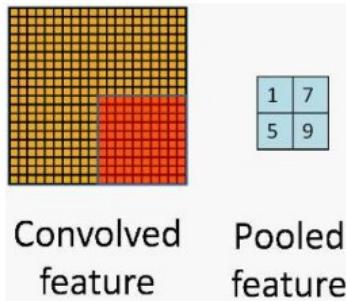
ادغام

بعد از به دست آوردن ویژگی‌ها توسط عملیات پیچش، می‌توان از آن‌ها برای دسته‌بندی استفاده کرد. به صورت تئوری می‌توان از یکی از دسته‌بندهای معمول برای دسته‌بندی توسط این ویژگی‌ها بهره جست. اما این کار از نظر محاسباتی بسیار چالش‌برانگیز است و تعداد ویژگی‌هایی که از هر یک از زیربخش‌های ورودی استخراج می‌شوند، بسیار سنگین بوده و احتمال بیش‌برازش^۱ را به مقدار زیادی بالا می‌برد.

همان‌طور که قبلاً اشاره شد، ویژگی‌هایی که از یک قسمت از تصویر استخراج می‌شود می‌توانند در قسمت‌های دیگر تصویر نیز به کار گرفته شود. برای بیان ویژگی‌های یک صحنه بزرگ، می‌توان از روش جمع‌کردن آماره‌های این ویژگی‌ها در محل‌های مختلف صحنه استفاده کرد. به عنوان مثال می‌توان میانگین (یا بیشینه) مقدار یک ویژگی را در یک محدوده خاص از تصویر به دست آورد. این آماره‌های خلاصه‌شده، ابعاد بسیار پایین‌تری نسبت به استفاده از تمام ویژگی‌های به دست آمده دارند. همچنین می‌توانند به دلیل کاهش احتمال بیش‌برازش، دقت الگوریتم را بهبود دهند. عملیات جمع‌آوری ویژگی‌ها را ادغام می‌نامند.

شکل (۳-۸) عملیات ادغام را بر روی چهار بخش بدون اشتراک نشان می‌دهد که در سمت چپ ویژگی‌هایی است که از پیچش به دست آمده و در سمت راست، خروجی ویژگی‌های ادغام‌شده را به نمایش درآمده است.

¹ Over-fitting



شکل (۸-۳) نمونه‌ای از عملیات ادغام ویژگی‌های پیچشی [۴۲]

تابع فعال‌سازی

در شبکه‌های عصبی، یک تابع فعال‌سازی مشخصی می‌کند ورودی‌های یک گره عصبی به ازای چه مقادیری دارای خروجی باشند. طی سال‌های گذشته انواعی از توابع فعال‌ساز بر حسب کاربرد تعریف و استفاده شده‌اند. در اینجا به ارائه توضیح در مورد سه مورد از آن‌ها بسنده می‌کنیم.

تابع فعال‌ساز سیگموئید^۱

تابع فعال‌ساز غیرخطی سیگموئید که بر روی یک نورون اعمال می‌شود ساختار ریاضیاتی به فرم زیر دارد:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (4-3)$$

این تابع فعال‌ساز، یک عدد با مقدار حقیقی را به عنوان ورودی دریافت می‌کند و آن را به بازه میان صفر و یک می‌برد. تابع فعال‌ساز سیگموئید به صورت سنتی در شبکه‌های عصبی بسیار پرکاربرد بوده است. دلیل جذابیت این تابع فعال‌ساز، نرخ تغییرات آن در نرخ فعال بودن^۲ یک نورون است که از صفر به معنی غیرفعال بودن تا یک به معنی فعال کامل گستردگ است. در شکل (۹-۳) نمودار خروجی این تابع به ازای ورودی در بازه [۰, ۱]- نمایش داده شده است. این تابع فعال‌ساز دارای مشکلاتی ذاتی جهت استفاده در شبکه‌های عمیق و یادگیری عمیق است [۴۳]. دو مشکل عمدۀ این تابع عبارت‌اند از:

❖ **محوشدگی گرادیان:** در صورتی که میزان فعال بودن یک نورون سیگموئید، در حدود صفر یا یک باشد، گرادیان در این ناحیه‌ها نزدیک به صفر خواهد بود. این امر در زمان پس انتشار^۳ مشکل‌ساز می‌شود. در واقع در زمان انتشار خطأ از انتهای شبکه به سمت ورودی، این گرادیان محلی (که

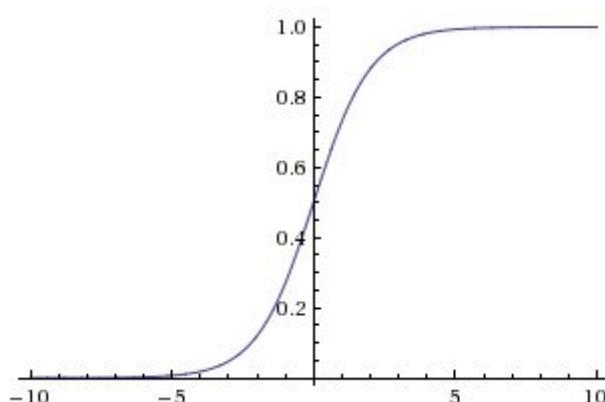
¹ Sigmoid

² Firing rate

³ Backpropagation

نژدیک به صفر هست) در گرادیان خروجی این دروازه^۱ ضرب خواهد شد. این ضرب منجر به محوشدگی گرادیان خواهد شد که باعث جلوگیری از انتشار خطا به لایه‌های پیشین می‌شود. برای جلوگیری از این رویداد، باید دقت مضاعفی در وزن دهی نخستین نورون‌های سیگموئید داشت. برای مثال در صورت وزن دهی اولیه با وزن‌های بسیار بزرگ، این نورون‌ها اشباع شده و گرادیان آن‌ها صفر خواهد بود و درنتیجه شبکه به سختی آموزش خواهد دید.

غیر صفر بودن میانگین خروجی: خروجی‌های به دست آمده از نورون‌های سیگموئید، به دلیل بازه خروجی این تابع، میانگینی غیر صفر دارند. این امر باعث دریافت دادگانی با میانگین غیر صفر در لایه‌های بعدی خواهد شد. به دلیل استفاده از روش کاهش گرادیان در آموزش شبکه‌های عصبی، در صورتی که در این شبکه‌ها دادگانی با مقادیر همیشه مثبت وجود داشته باشد، (برای مثال $x > 0$ در $f = w^T x + b$) آنگاه گرادیان بر روی وزن w در زمان پس انتشار، تمام مثبت یا تمام منفی خواهد شد (بسته به گرادیان تمام عبارت f). این امر می‌تواند منجر به تغییرات کجومعوج^۲ در تغییرات به روزرسانی گرادیان برای وزن‌ها شود. با این وجود، اگر گرادیان‌ها در یک دسته^۳ دادگان محاسبه و اعمال شود، به روزرسانی نهایی وزن‌ها می‌تواند علامت (مثبت یا منفی) متغیری داشته باشد که تا حدی کاهش‌دهنده این مشکلات است. با وجود اینکه این مورد، یک مشکل محسوب می‌شود، اما تأثیر منفی کمتری در مقایسه با مورد پیشین (محوشدگی گرادیان) دارد.



شکل (۹-۳) خروجی تابع فعال‌ساز سیگموئید

تابع فعال‌سازی Tanh

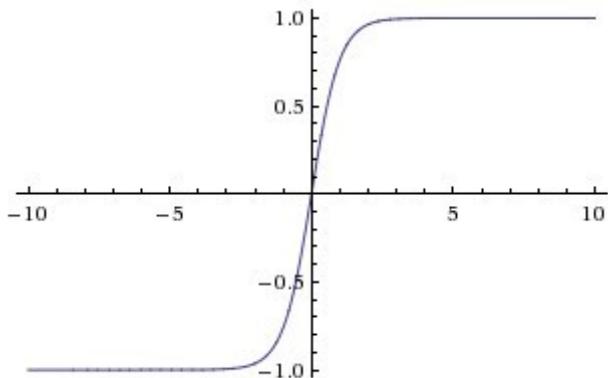
¹ Gate

² Zig-zag

³ batch

تابع فعال‌ساز غیرخطی Tanh، اعداد ورودی از جنس حقیقی را به بازه $[-1, 1]$ ^۱ می‌برد. این تابع، همانند تابع Sigmoid، مشکل اشباع فعال‌سازی (درنتیجه محوشدگی گرادیان) را دارد. اما مشکل غیر صفر بودن میانگین را ندارد. تابع Tanh در واقع، حالت خاصی از تابع Sigmoid است. در رابطه زیر تابع Tanh مشاهده می‌شود. همچنین در شکل (۱۰-۳) خروجی این تابع به ازای بازه ورودی $[-10, 10]$ به تصویر کشیده شده است.

$$\tanh(x) = 2\sigma(2x) - 1 \quad (5-3)$$



شکل (۱۰-۳) تابع فعال‌ساز Tanh

تابع فعال‌ساز ReLU

تابع فعال‌ساز ReLU^۱ (به معنای واحد خطی تصحیح شده) در یادگیری ژرف بسیار پرکاربرد است. این تابع در رابطه زیر نمایش داده شده است:

$$f(x) = \max(0, x) \quad (6-3)$$

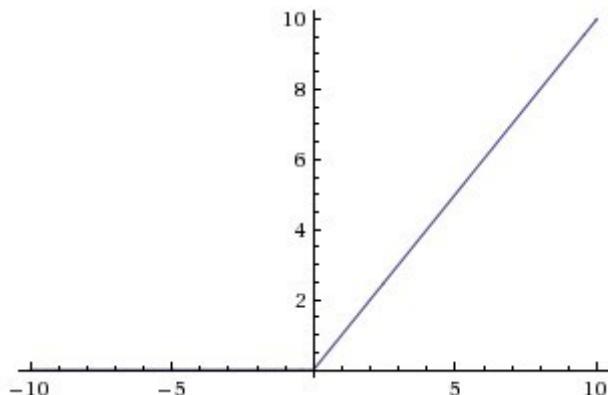
شکل خروجی این تابع در شکل (۱۱-۲) نمایش داده شده است. این تابع دارای مزایای زیر نسبت به تابع سیگموئید و Tanh است:

^۱ Rectified Linear Unit

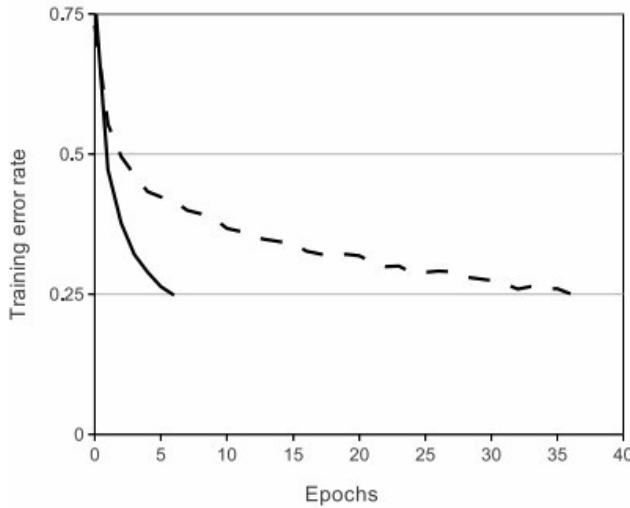
- این تابع به دلیل ساختار خطی غیر اشباع شونده‌اش، سرعت همگرایی روش کاهش گرادیان را بهشت افزایش می‌دهد و برای کاربردهای مبتنی بر یادگیری ژرف بسیار مناسب است [۴۳]. در شکل (۱۱-۳) مقایسه بازدهی آموزشی این تابع با تابع Tanh نمایش داده شده است.

- برخلاف توابع سیگموئید و Tanh که شامل عملیات هزینه‌بر ریاضی می‌شوند، تابع ReLU تنها به محاسبه نورون‌های بیشتر از صفر می‌پردازد.

این تابع دارای یک مشکل اساسی است. تابع ReLU در زمان آموزش، احتمال از کارافتادن دارد که اصطلاحاً به «مرگ» ReLU معروف است. در صورت حرکت یک گرادیان بزرگ از این تابع، می‌تواند وزن‌های نورون موردنظر را به گونه‌ای تغییر دهد که نورون هیچ‌گاه دوباره فعال نشود. در صورت روی دادن این مشکل، گرادیان عبودی از این واحد، برای همیشه صفر خواهد بود. به همین دلیل به این مسئله «مرگ» غیرقابل احیا واحدهای ReLU گفته می‌شود. یکی از عوامل تأثیرگذار بر این مسئله، نرخ یادگیری است. در صورت بالا بودن نرخ یادگیری، احتمال روی دادن این مشکل بسیار زیاد می‌شود. در صورت تنظیم مناسب نرخ یادگیری، این مشکل بهندرت پیش می‌آید.



شکل (۱۱-۳) خروجی تابع فعال‌ساز ReLU



شکل (۱۲-۳) تصویری از مقاله [۴۳] برای نمایش بهبود شش برابر همگرایی با ReLU (خط ساده) در برابر همگرایی با Tanh (خطچین)

۳-۳-۱- مدل‌های گرافیکی احتمالاتی^۱

مدل‌های گرافیکی ترکیبی از تئوری احتمال و تئوری گراف می‌باشند. در این روش ابزارهای طبیعی فراهم می‌گردد که از طریق آن‌ها می‌توان مسائل مربوط به ریاضی کاربردی و مهندسی که پیچیده و غیرقطعی هستند را حل کرد و علاوه بر آن نقش مهم و رو به افزونی در مورد الگوریتم‌های بادگیری ماشین دارند. مدل گرافیکی یک ساختار رسمی ریاضیاتی را فراهم می‌کند که امکان درک انواع مختلفی از شبکه‌ها محاسباتی را به وجود می‌آورد.

ایده اصلی در طراحی مدل گرافیکی استفاده از ساختار پیمانه‌ای^۲ است از تئوری گرافی برای ایجاد یک رابط مناسب استفاده می‌شود که می‌تواند تعامل بالای مجموعه داده‌ای را مدل نماید. تئوری ریاضی (احتمال) به عنوان ارتباط‌دهنده‌ی بین پیمانه‌ها است و سیستم درمجموع به صورت یکپارچه است. مهندسی سیستم‌ها، تئوری اطلاعات، تشخیص الگو موارد ویژه‌ای از فرم‌های عمومی مدل گرافیکی است.

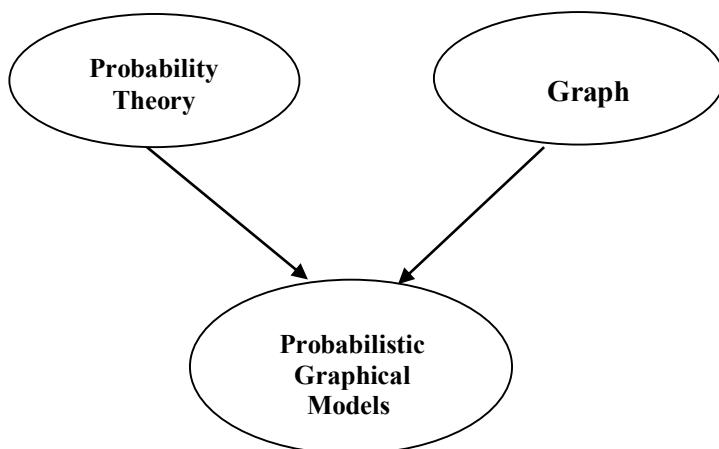
مدل‌های گرافیکی زیرساختی را برای معرفی مدل‌هایی که در آن تعدادی متغیر باهم تعامل می‌کنند را ایجاد می‌کنند. هر گره در گراف یک متغیر تصادفی را معرفی می‌کند و نوعی یال در گراف وابستگی کیفی

¹ Probabilistic Graphical Models

² module

بین متغیرها را نشان می‌دهد. عدم وجود این یال به معنای نبود وابستگی این متغیر به متغیرهای دیگر است. وابستگی مقداری بین گره‌های متصل از طریق توزیع شرطی پارامتری شده، بیان می‌شود. الگوی یال‌ها وتابع پتانسیل یک توزیع پیوسته را روی همه متغیرهای گراف نشان می‌دهد. الگوی یال‌ها ساختار گراف را نشان می‌دهد. از این‌رو یک مدل گرافیکی، روشی برای معرفی احتمال رابطه‌ای بین متغیرهای تصادفی استفاده می‌شود.

تصویر زیر نشان می‌دهد مدل گرافیکی از ترکیب تئوری گراف و تئوری احتمال ایجاد شده است.



شکل (۱۳-۳) مدل گرافیکی احتمالاتی از دو بخش نظریه احتمال و گراف تشکیل شده است

انواع مختلفی از مدل‌های گرافی احتمالی وجود دارد، اما ویژگی مشترک همگی این است که با استفاده از یک گراف، یک خانواده از توزیع‌های احتمالی را مشخص می‌کنند. انواع مختلف این مدل‌ها در ساختار گراف و فرض‌های استقلال شرطی که در گراف در نظر گرفته شده، با یکدیگر تفاوت دارند. با داشتن یک مدل گرافی احتمالاتی می‌توان آن را به صورت یک فیلتر برای توزیع‌های احتمالی در نظر گرفت که فقط توزیع‌هایی از آن عبور می‌کنند که همه شرط‌های استقلالی که در گراف در نظر گرفته شده را ارضا کنند.^[۴۴] در نتیجه یک مدل گرافی احتمالی فقط یک توزیع را مشخص نمی‌کند بلکه یک خانواده از توزیع‌های احتمالی را مشخص می‌کند.

معمولًا در استفاده از مدل‌های گرافیکی لازم است تا عملیات زیر انجام گیرد:

۱- تعریف مدل:

این مرحله شامل تعیین متغیرهای تصادفی به عنوان گره‌های مدل گرافیکی، تعیین توزیع احتمال گره‌ها و تعیین وابستگی بین گره‌ها به صورت یال‌ها در مدل است.

۲- ایجاد الگوریتم استنتاج :

در این مرحله یک الگوریتم برای استنتاج ایجاد می‌گردد.

۳- استفاده از الگوریتم استنتاج

یک مدل گرافیکی به صورت $G=(X,E)$ تعریف می‌شود که گره‌های X برای نمایش متغیرهای تصادفی به کار می‌روند (یا به عبارت کلی تر برای نمایش مجموعه‌ای از متغیرها است) که از تابع توزیع احتمال $p(X)$ پیروی می‌کنند. یال‌های E برای تعیین توزیع احتمال وابستگی بین گره‌ها به کار می‌رود، به این صورت که وجود یک یال بیان‌کننده وابستگی بین گره‌ها بوده و نبود یال بین گراف‌ها بیان‌کننده عدم وابستگی بین گره‌ها است.

نکته موردتوجه در این قسمت این است که در مدل گرافیکی توزیع پیوسته احتمال، برای گره‌های مختلف چگونه انجام می‌گیرد و این توزیع پیوسته چگونه محاسبه می‌گردد؟ پاسخ این است که برای هر گره با استفاده از قانون احتمال شرطی، مقدار احتمال محاسبه شده و به هر گره نسبت داده می‌شود.

با استفاده از حاصل‌ضرب احتمالات داخلی (محلی)، احتمالات سراسری را تولید می‌کند. در انجام محاسبات از دو قانون احتمال زیر استفاده می‌گردد.

قانون اول معروف به قانون مجموع احتمال است:

$$P(A) = \sum_B P(A, B)$$

قانون دوم نیز معروف به قانون حاصل‌ضرب است:

$$P(A, B) = P(B|A)P(A)$$

از جمله مزایای مدل گرافیکی می‌توان به موارد زیر اشاره کرد:

- ❖ استنتاج و یادگیری به صورت وابسته به هم و با همدیگر انجام می‌شوند.
- ❖ یادگیری با ناظر و بدون ناظر به صورت یکپارچه انجام می‌شود.
- ❖ در مقابل ویژگی‌های بدون داده رفتار غیرقابل پیش‌بینی ندارند.
- ❖ تمرکز روی استقلال شرطی و عملیات محاسباتی.
- ❖ در صورت لزوم قابلیت تفسیر دارد.

نحوه‌ی نمایش

همان‌طور که بیان شد مدل‌های گرافیکی احتمالی ، گراف‌هایی هستند که گره‌ها متغیرهای تصادفی را نشان می‌دهند و یال‌ها فرض شرطی مستقل را معرفی می‌کنند . از این‌رو یک نمایش پیوسته از توزیع احتمال پیوسته را نمایش می‌دهند . به عنوان مثال یک نمایش اتمی از $P(X_1 \dots X_n)$ با استفاده از N متغیر تصادفی باین‌ری ، به $O^{(2^n)}$ پارامتر نیاز دارند و به همین ترتیب یک مدل گرافیکی ممکن است به تعداد نمایی پارامتر داشته باشد .

به‌طور کلی می‌توان گفت دو مدل اساسی از مدل گرافیکی ارائه شده است : نمایش گرافی جهت‌دار(بیزی) و نمایش گرافی بدون جهت (مارکوف).

شبکه‌های بیزی با گراف‌های جهت‌دار بازنمایی می‌شوند و بنابراین مدل‌های گرافی جهت‌دار نامیده می‌شوند. بازنمایی شبکه‌های مارکوف نیز با استفاده از گراف‌های بدون جهت بوده و مدل‌های گرافی بدون جهت نامیده می‌شوند. همچنین ممکن است مدل‌های ترکیبی که شامل هر دو بازنمایی جهت‌دار و بدون جهت هستند داشته باشیم که به اندازه دو دسته قبلی متداول نیستند. مقالات زیادی در دهه‌های گذشته ارائه شده است که استفاده از مدل‌های گرافی برای تجزیه و تحلیل داده‌های متوالی را توصیه کرده است. بیشتر روش‌های موجود این رده احتمال و نظریه گراف را برای پیدا کردن ساختار در داده‌های ترکیبی ترکیب نموده‌اند. این روش را می‌توان به‌طور گسترده به دو زیرمجموعه مدل‌های گرافی جهت‌دار و مدل‌های گرافی بدون جهت تقسیم کرد. از جمله مدل‌های معروف دسته اول شامل مدل مارکوف پنهان^۱ و شبکه‌های بیزی^۲ هستند. میدان‌های تصادفی مارکوف^۳ و میدان‌های تصادفی شرطی^۴ نیز از جمله مدل‌های متعلق به دسته دوم هستند. ساده‌ترین مورد از مدل گرافی جهت‌دار مدل مارکوف مخفی^۵ است که در آن مشاهدات در حالت جاری به حالت‌های قبلی وابستگی دارد. مشاهدات می‌توانند به عنوان نمادهای گسسته یا توزیع پیوسته نمایش داده شوند. در تشخیص رویدادها و فعالیت‌های پیچیده، مدل گرافی جهت‌دار توسط گراف بدون دور نشان داده می‌شود که می‌توانند روابط فضایی حالت را بین رویدادهای سطح پایین و یا زیر رویدادها را مدل کند.

¹ Hidden Markov Model

² Basian networks

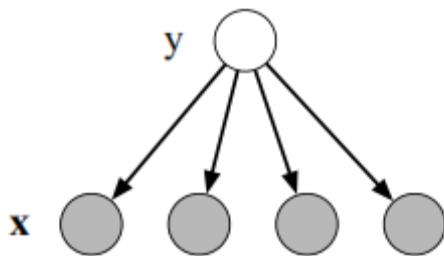
³ Markov Random Field

⁴ Conditional Random Field

⁵ Hidden Markov Model

شبکه‌های بیزی

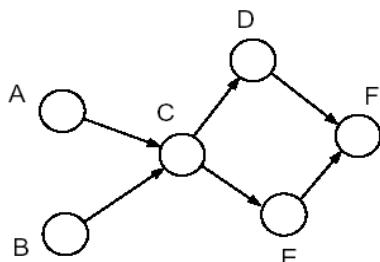
شبکه‌های بیزی با گراف‌های جهت‌دار بدون حلقه بازنمایی می‌شوند. رأس‌های گراف نمایانگر متغیرهای تصادفی و یال‌ها نیز به صورت شهودی نشان‌دهنده تأثیر مستقیم یک متغیر روی متغیر دیگر است. این گراف را می‌توان به صورت یک ساختار داده برای بازنمایی توزیع احتمال توأم در نظر گرفت. به عنوان مثال در شکل زیر شبکه مربوط به دسته‌بند بیز ساده^[۴۵] از جمله مدل‌های گرافی جهت‌دار قبل مشاهده است.



شکل (۱۴-۳) یک مثال از دسته‌بند بیز ساده

هر متغیر تصادفی در شبکه برای خودش یک توزیع احتمال شرطی و یا یک مدل احتمال محلی دارد که آن را توزیع احتمال شرطی یا CPD^۲ می‌نامیم. برای هر متغیر X در شبکه، با داشتن والدین آن متغیر در گراف، CPD برابر است با احتمال شرطی متغیر X به شرط داشتن والدین X را می‌توان به روش‌های مختلفی بازنمایی کرد، یکی از این روش‌های استفاده از جدول است که به ازای هر مقدار ممکن برای والدین متغیر X دارای یک ردیف است که در آن احتمال مقادیر مختلف X داده شده است.

به عنوان مثال در گراف زیر با استفاده از این روش یک توزیع احتمال P(A,B,C,D,E,F) به گراف نسبت داده می‌شود و تمام محاسبات بر اساس این توزیع انجام می‌گیرد.



$$P(F|A,B) = \frac{\sum_C \sum_D \sum_E P(A, B, C, D, E, F)}{\sum_C \sum_D \sum_E \sum_F P(A, B, C, D, E, F)}$$

شکل (۱۵-۳) مثالی از یک گراف جهت‌دار و توزیع احتمال مربوط به آن

در مدل گرافیکی جهت‌دار (شبکه بیزین) یک یال از A به B به این معنی است که A سبب ایجاد B

¹ Naive Bayes

² Conditional Probability Distribution

است.

همان‌طور که در مقدمه اشاره شد یکی از ساده‌ترین انواع مدل‌های گرافیکی احتمالاتی، مدل مخفی مارکوف است. یک نقطه ضعف از مدل مخفی مارکوف عدم توانایی آن در مدل کردن رابطه علی است. این مشکل توسط نوع متفاوت از مدل گرافی شبکه‌های جهت‌دار به نام بیزی که با اختصار آن را با^۱ BN نمایش می‌دهیم، حل می‌گردد. شبکه‌های بیزی قادر به مدل‌سازی مؤثر علیت با استفاده از استقلال شرطی بین حالت‌ها می‌باشند. این روش با اعمال فاکتور باعث تسهیل معنایی و محاسباتی در فضای حالت می‌شود. شبکه‌های بیزی به‌طور ضمنی نمی‌توانند اطلاعات زمانی بین گره‌ها یا حالات مختلف در مدل ماشین حالت محدود را مدل کنند. شبکه‌های بیزی پویا (DHN) با استفاده از اصول فاکتورگیری که در شبکه‌های بیزی وجود دارد، روابط زمانی را به‌صورت بهتری مدل می‌کنند. شبکه‌های بیزی و مدل‌های مخفی مارکوف و انواع آن‌ها از لحاظ فلسفی در گروه مدل‌های مولد قرار می‌گیرند. به دلیل ماهیت مولد مدل، توزیع از روی مشاهدات با توجه به حالت یاد گرفته می‌شود. با این حال، در زمان استنتاج و یا دسته‌بندی، تنها مشاهدات در دسترسی می‌باشند. از این‌رو از نظر شهودی، شرط گذاشتن بر روی مشاهدات بهتر از شرط گذاشتن بر روی حالت‌ها است. این موضوع انگیزه‌ای برای محققان شد تا مدل‌سازی حوادث پیچیده با استفاده از مدل‌های گرافی بدون جهت را مورد مطالعه قرار دهند.^[۴۶]

تئوری بیز

فرض کنید که فضای فرضیه H و مجموعه مثال‌های آموزش D موجود باشند. مقادیر احتمال زیر را تعریف می‌کنیم:

۱. $P(h)$ = احتمال پیشین یا احتمال اولیه‌ای که فرضیه h قبل از مشاهده مثال آموزشی D داشته

است. اگر چنین احتمالی موجود نباشد می‌توان به تمامی فرضیه‌ها احتمال یکسانی نسبت داد.

۲. $P(D)$ = احتمال داده‌های آموزشی D که مشاهده می‌شود.

۳. $P(D|h)$ = احتمال مشاهده داده آموزشی D به فرض آنکه فرضیه h صادق باشد.

در یادگیری ماشین علاقه‌مند به دانستن $P(h|D)$ یعنی احتمال اینکه با مشاهده داده آموزشی D فرضیه h صادق باشد، هستیم. این رابطه احتمال ثانویه^۲ نامیده می‌شود. می‌دانیم که احتمال اولیه مستقل از داده آموزشی است ولی احتمال ثانویه تأثیر داده آموزشی را منعکس می‌کند.

¹ Bayesian Network

² posterior probability

ویژگی‌ها یادگیری بیزی

مشاهده هر مثال می‌تواند به صورت جزئی باعث افزایش و یا کاهش احتمال درست بودن یک فرضیه گردد. برای به دست آوردن احتمال یک فرضیه می‌توان دانش قبلی را با مثال مشاهده شده ترکیب کرد. این دانش قبلی به دو طریق به دست می‌آید:

۱. احتمال پیشین برای هر فرضیه موجود باشد
 ۲. برای داده مشاهده شده توزیع احتمال هر فرضیه ممکن موجود باشد
- روش‌های بیزی فرضیه‌هایی ارائه می‌دهند که قادر به پیش‌بینی احتمالی هستند. حتی در مواردی که روش‌های بیزی قابل محاسبه نباشند، می‌توان از آن‌ها به عنوان معیاری برای ارزیابی روش‌های دیگر استفاده کرد.

نیاز به دانش اولیه در نظریه بیز مستلزم محاسبه تعداد زیادی مقادیر احتمال است. وقتی که این اطلاعات موجود نباشند اغلب ناگزیر به تخمین زدن آن هستیم. برای این کار از اطلاعات زمینه، داده‌ایی که قبلاً جمع‌آوری شده‌اند، و فرضیاتی در مورد توزیع احتمال استفاده می‌شود.

تئوری بیز

در یادگیری ماشین معمولاً در فضای فرضیه H به دنبال بهترین فرضیه‌ای هستیم که در مورد داده‌های آموزشی D صدق کند. یکراحت تعیین بهترین فرضیه، این است که به دنبال محتمل‌ترین فرضیه‌ای باشیم که با داشتن داده‌های آموزشی D و احتمال قبلی در مورد فرضیه‌های مختلف می‌توان انتظار داشت.

- تئوری بیز چنین راه حلی را ارائه می‌دهد. این روش راه حل مستقیمی است که نیازی به جستجو ندارد.
- سنگ بنای یادگیری بیزی را تئوری بیز تشکیل می‌دهد. این تئوری امکان محاسبه احتمال ثانویه را بر مبنای احتمالات اولیه می‌دهد:

$$P(h | D) = \frac{P(D | h)P(h)}{P(D)} \quad (7-3)$$

همان‌طور که در رابطه (۴-۳) مشاهده می‌شود با افزایش $P(h|D)$ کاهش می‌یابد. زیرا هر چه احتمال مشاهده D مستقل از h بیشتر باشد به این معنا خواهد بود که D شواهد کمتری در حمایت از h در بردارد.

شبکه‌های مارکوف

دومین دسته از مدل‌های گرافیکی احتمالاتی شبکه‌های مارکوف یا میدان تصادفی مارکوف نامیده می‌شود. اساس این مدل‌ها بدون جهت بودنشان است. این مدل‌ها برای مدل‌سازی انواع مختلفی از پدیده‌ها قابل استفاده هستند که در آن‌ها نمی‌توان جهتی را بین متغیرهای تصادفی مختلف متصور شد. علاوه بر این، مدل‌های بدون جهت هم از لحاظ ساختار استقلال شرطی و هم از لحاظ استنتاج نسبت به مدل‌های جهت‌دار ساده‌تر هستند.

یک نوع از بازنمایی برای پیاده‌سازی این نوع از دیدگاه استفاده از گراف بدون جهت است. همانند شبکه‌های بیزی رأس‌های گراف نشان‌دهنده متغیرها هستند و یال‌های گراف نمایانگر ارتباط بین متغیرها هستند.

موضوع دیگری که در این ارتباط مطرح است چگونگی بیان این مدل‌ها در قالب پارامترهای مختلف است. ساختار گراف ویژگی‌های کیفی توزیع احتمالی را توصیف می‌کند. برای بازنمایی این توزیع احتمالی، باید ساختار گراف را به تعدادی پارامتر ارتباط دهیم، مشابه تعریف CPD برای مدل‌های جهت‌دار که بالاتر به آن اشاره شد، با این تفاوت که در شبکه‌های مارکوف فاکتورها متناظر با احتمال یا احتمال شرطی نیستند. برای تبدیل این فاکتورها به احتمال نیاز به نرمال کردنشان با استفاده ازتابع پتانسیل هست.^[۴۶]

تعریف:

اگر H را یک شبکه مارکوف در نظر بگیریم، توزیع P_H را این شبکه مارکوف را توصیف می‌کند به شرطی که:

- مجموعه‌ای از زیرمجموعه‌های D_1, D_m, \dots, D_n وجود داشته باشد به‌طوری‌که هر D_i یک زیر گراف کامل از H باشد.
- فاکتورهای $\pi_1[D_1], \dots, \pi_m[D_m]$ را داشته باشیم به طوری‌که :

$$P_H(X_1, \dots, X_n) = 1/Z P'(X_1, \dots, X_n) \quad (8-3)$$

که در آن:

$$P'(X_1, \dots, X_n) = \pi_1[D_1] * \dots * \pi_m[D_m] \quad (9-3)$$

نرمال نشده است. هم‌چنین داریم:

$$Z = \sum_{X_1, X_2, \dots, X_n} P'(X_1, \dots, X_n) \quad (10-3)$$

Z یک مقدار ثابت برای نرمال کردن فاکتورها است. این عبارت تابع پتانسیل نامیده می‌شود. یک توزیع احتمال P که روی شبکه H تعریف می‌شود را توزیع گیبس^۱ روی H می‌نامیم. (این نام‌گذاری در فیزیک آماری ریشه دارد.) در شبکه‌های مارکوف تنها قید روی پارامترهای موجود در فاکتورها غیر منفی بودن آن‌ها است.

با توجه به اینکه هر زیر گراف کامل یک کلیک^۲ است، می‌توان پارامتر کردن مدل را با تعریف فاکتور تنها برای کلیک‌ها ساده کرد. به عبارت دیگر، می‌توان فاکتوری برای زیر کلیک‌ها تعریف نکرد. این فاکتورها پتانسیل کلیک نامیده می‌شوند.

استقلال متغیرها در شبکه‌های مارکوف

در شبکه‌های مارکوف، ساختار گراف را می‌توان تعدادی فرض استقلال بین متغیرها در نظر گرفت. به صورت شهودی، در این شبکه‌ها تأثیر احتمالات از طریق یال‌های منتقل می‌شود. دو دسته فرض استقلال بین متغیرها می‌توان در نظر گرفت، ویژگی‌های مارکوف محلی و ویژگی‌های مارکوف سراسری. ویژگی‌های محلی مارکوف مربوط به یک متغیر و متغیرهای همسایه آن می‌شود، به این صورت که می‌توان با شرط داشتن متغیرهای همسایه یک متغیر از تأثیر سایر متغیرها روی آن متغیر جلوگیری کرد. به عبارت دیگر طبق فرض مارکوف یک متغیر تصادفی در گراف از بقیه متغیرها مستقل خواهد بود در صورتی که مقدار همسایه‌های آن را داشته باشیم.

مدل‌های تولیدی^۳ و جداساز^۴

یک تفاوت مهم بین مدل‌ها بیز ساده(NB) و رگرسیون لجستیک^۵ (lg) این است که NB مدلی تولیدی بوده یعنی مبتنی بر یک توزیع احتمال توأم است. در حالی که lg یک مدل جداساز است، به این معنی که بر اساس یک توزیع احتمال شرطی است. حال به تفاوت بین مدل‌های جداساز و تولیدی و مزیت‌های مدل‌های تولیدی در بسیاری از موارد پرداخته خواهد شد. برای ملموس‌تر بودن مثال‌ها از مدل‌های lg و NB خواهند بود ولی بحث کلی مقایسه بین مدل‌های جداساز و تولیدی است.^[۴۵]

تفاوت اصلی بین این دو نوع مدل این است که توزیع شرطی حاوی مدل $(x)p$ نیست، به دلیل اینکه این اطلاعات برای دسته‌بندی نیاز نیست. مشکل اصلی در مدل کردن $(x)p$ این است که معمولاً شامل تعداد

¹ Gibbs Distribution

² clique

³ generative

⁴ discriminative

⁵ logistic regression

زیادی ویژگی به هم وابسته است که برای مدل کردن دشوار است. برای استفاده از ویژگی‌های مستقل از هم در یک مدل تولیدی، ما دو راه داریم: بهبود مدل برای بازنمایی وابستگی‌های بین ورودی‌ها، و یا در نظر گرفتن فرض‌هایی برای ساده‌سازی مانند فرض موجود در مدل بیز ساده انجام دیدگاه اول یعنی بهبود مدل عموماً دشوار است. دیدگاه دوم، یعنی استفاده از فرض استقلال بین ورودی‌ها، ممکن است که عملکرد را دچار مشکل کند. برای مثال با اینکه دسته‌بند NB در دسته‌بندی متون عملکرد خوبی دارد، اما در کل دقت آن پایین‌تر از رگرسیون لجستیک است.

مزیت اصلی مدل‌های جداساز این است که برای حالتی که ویژگی‌های دارای همپوشانی داریم بهتر عمل می‌کنند.

ساختار مدل گرافی [۴۶]

یکی دیگر از مسائلی که در هنگام استفاده از مدل‌های گرافی باید مشخص شود انتخاب ساختار مناسب مدل گرافی است. به‌طور کلی در ساخت مدل‌های گرافی دو استراتژی وجود دارد ۱- آموزشی خودکار ساختار مدل گرافی (یادگیری ساختار) بر اساس تعدادی قید ۲- ساخت دستی ساختار مدل بر اساس دانش قبلی که انسان از مسئله دارد. هرچند یادگیری خودکار ساختار می‌تواند تطبیق بیشتری با مسئله داشته باشد و نتایج بهتری را در برداشته باشد ولی یادگیری چنین ساختاری حتی برای مدل‌های گرافی ساده مانند شبکه‌های بیزی بسیار دشوار است. در اینجا تمرکز ما بر روی ساخت ساختار مدل به صورت دستی است. در صورتی که رابطه به صورت علی و یا یک طرفه باشد از پیوندهای جهت‌دار و در صورتی که رابطه متقابل یا دوطرفه باشد از پیوندهای بدون جهت استفاده می‌کنیم. در حالاتی که بتوان از هر دو نوع ارتباط استفاده کرد مدلی را انتخاب می‌کنیم که به ساده شدن مدل کلی گراف منجر شود.

به دست آوردن پارامترهای مدل‌های گرافی

هنگامی که از روی مدل ساخته شده توزیع احتمال توام (JPD)^۱ به دست می‌آوریم تعدادی پارامتر مجھول در JPD باقی می‌مانند که می‌بایست به دست آیند. به این فرایند پارامتر سازی می‌گویند. به طور کلی در صورتی که پیوندها بدون جهت باشند، پارامترها با استفاده از توابع پتانسیل به دست آمده ولی در صورتی که پیوندها جهت دار باشند، پارامترها با استفاده از احتمالات شرطی محلی به دست می‌آیند.

یادگیری

به‌طور کلی یادگیری پارامترها در مدل‌های گرافی بدون جهت مشکل‌تر از یادگیری پارامترها در مدل‌های

¹ Joint Probability Distribution

گرافی جهتدار است. به همین جهت در یادگیری پارامترهای مدل‌های گرافی بدون جهت مسئله را با تغییر یا تقریب تابع هدف ساده می‌کنند و از به دست آوردن مقدار دقیق تابع هدف یا مشتق آن اجتناب می‌کنند. تلاش‌های زیادی به منظور یادگیری پارامترها در مدل‌های گرافی جهتدار و بدون جهت انجام شده است ولی کارهای زیادی در رابطه با یادگیری مدل‌های گرافی ترکیبی صورت نگرفته است. یکی از راه‌های عمومی به منظور یادگیری پارامترها در این مسئله استفاده از روش MLE¹ است. فرض کنید تعداد K عدد داده آموزش i.i.d. به صورت داریم که در آن X بیانگر مقدار همه متغیرهای تصادفی در امین k نمونه است. هدف MLE بیشینه کردن لگاریتم احتمال پارامترها است.

استنتاج

در مدل گرافیکی برخلاف شبکه‌های دیگر که وزن یک تأثیر محلی دارد، در الگوریتم استنتاج یک تأثیر سراسری روی متغیرها دارد که برای حصول این نتیجه از یک توزیع احتمال پیوسته برای شبکه استفاده شده است. استنتاج یک عمل احتمالی است که با استفاده از توزیع‌های احتمال مقادیر احتمال کناری و شرطی را به دست می‌آورد.

هدف اصلی از استنتاج؛ تخمین مقادیر متغیرهای مخفی است که بر اساس مقادیر داده شده در متغیرهای قابل مشاهده، محاسبه می‌شوند.

هر دو نوع مدل‌های جهتدار و غیر جهتدار یک توزیع احتمالی روی یک مجموعه متغیر تعریف می‌کنند. پاسخ برخی پرس‌وجوهای² ممکن را می‌توان با استفاده از توزیع احتمال توأم پاسخ داد. یکی از متداول‌ترین انواع پرس‌وجوه، احتمال شرطی است که به صورت $P(Y | E = e)$ نوشته می‌شود. این پرس‌وجو شامل دو قسمت است: مشاهده که با E مشخص شده و متغیر تصادفی که مورد پرسش قرار گرفته که در عبارت فوق با Y مشخص شده است. در اینجا عبارت خواسته شده، حاصل تقسیم احتمال توأم Y و e بر مجموع احتمالات Y به ازای $e = e$ خواهد بود [۴۶].

گراف عامل³ یکی از راه‌هایی است که می‌توان با استفاده از آن در مدل‌های گرافی استنتاج انجام داد ولی در ابتدا گراف را در هر فرمی که است باید به شکل گراف عامل تبدیل کرد و سپس با روش‌های موجود استنتاج را انجام داد. تبدیل مدل‌های گرافی جهتدار و بدون جهت به گراف عامل به سادگی امکان‌پذیر است. گراف عامل در مدل‌های گرافی، بین گره‌های موجود که هر کدام نشان‌دهنده یک متغیر می‌باشند یک گره اضافه می‌کند که معادل تابع محلی است، پس از آن بین گره‌های اصلی و گره‌های اضافه شده پیوندهای

¹ Maximum Likelihood Estimation

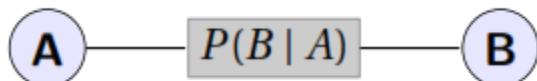
² query

³ Factor Graph

بدون جهت اضافه می‌شود تا گراف عامل به دست آید. درنتیجه تابع کلی گراف یا استفاده از توابع محلی به دست می‌آید. در مدل‌های گرافی جهت‌دار مانند BN توابع گره‌های عامل احتمالات شرطی بین متغیرها می‌باشند و گره‌های عادی گراف عامل متغیرهای توزیع هستند. برای مثال BN شکل زیر را در نظر بگیرید [۴۶].



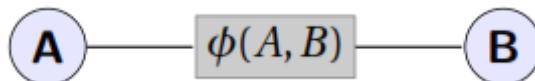
گراف عامل آن به شکل زیر خواهد بود:



مقدار $P(B | A)$ در گره عامل قرار می‌گیرد. در مدل‌های گرافی بدون جهت، نیز به همین صورت عمل می‌شود با این تفاوت که در گره‌های عامل مقدار تابع پتانسیل بین دو گره قرار می‌گیرد. برای مثال گراف زیر را در نظر بگیرید.



گراف عامل آن به شکل زیر است که همان‌طور که مشاهده می‌کنید مقدار $\phi(A, B)$ در گره عامل آن قرار گرفته است.



گره عامل در این گراف معادل با تابع پتانسیل در تابع توزیع توان است. بنا بر آنچه بیان شد می‌توان گراف عامل را در مدل‌های گرافی ترکیبی به دست آورد. در مدل‌های گرافی ترکیبی توزیع احتمال توان حاصل ضرب دو عامل تابع پتانسیل و احتمالات شرطی است. با داشتن این احتمال توزیع توان برای هر زیر گراف ما می‌توانیم با اعمال قوانینی که در بالا به آن اشاره شد مدل گرافی توان را به نمایش گراف عامل تبدیل کنیم. در این گراف عامل گره‌های متغیرها همان متغیرهای تصادفی هستند که در مدل گرافی مرکب حضور داشتنند. بعلاوه گره‌های عامل به دلیل نمایش تابع پتانسیل یا احتمال توان به این گراف اضافه می‌شوند. با روشنی که بیان شد، می‌توان گراف عامل را به دست آورد و پس از آن با روش‌های مختلفی استنتاج را روی آن انجام داد. روش‌های جمع ضرب‌ها¹ و بیشینه ضرب‌ها² دو روش معروف به منظور استنتاج هستند. روش جمع ضرب‌ها بر اساس سازوکار تبادل پیام عمل می‌کند. دو نوع پیام متصور است: یکی پیام‌هایی که از متغیرها به گره‌های عامل ارسال می‌شود و دیگری پیام از گره عامل به متغیر، در صورتی که پیام‌ها به درستی

¹ Sum-product

² Max-product

مقداردهی اولیه شوند، پس از تعدادی به روزرسانی گراف عامل همگرا خواهد شد. احتمال‌های حاشیه^۱ ای یا استفاده از این پیام‌ها قابل محاسبه می‌باشند. در صورتی که گراف به شکل درخت باشد، احتمال‌های حاشیه‌ای به صورت دقیق استنتاج خواهد شد. احتمال‌های حاشیه‌ای به صورت زیر قابل محاسبه است.

$$P(X_i) = \sum_{X \setminus X_i} P(X)$$

که در رابطه بالا X نشان‌دهنده تمامی متغیرها است و $X_i \in X$ هر کدام از متغیرها است. همان‌طور که مشخص است جمع بر روی تمامی متغیرها به غیر از متغیری که می‌خواهیم توزیع آن را به دست آوریم انجام می‌شود. با داشتن توزیع حاشیه‌ای هر متغیر می‌توان مقدار بهینه حالت آن را یا استفاده از روش MAP^۲ به دست آورد. روش دیگر استفاده از الگوریتم بیشینه ضرب‌ها است که در آن مقدار متغیرها به نحوی تعیین می‌شود که ضرب متغیرها یا به عبارت دیگر توزیع احتمال توأم مقدار بیشینه را داشته باشد.

$$X_i^* = \operatorname{argmax}_{x_i} P(X_i)$$

۳-۲-۳- علت انتخاب روش

همان‌طور که پیش‌ازین بیان شد در سال‌های اخیر با افزایش قدرت محاسباتی رایانه‌ها و امکان ایجاد و آموزش شبکه‌های عمیق با پارامترهای بسیار زیاد، یادگیری عمیق عملکرد مطلوبی را در حوزه‌های مختلف یادگیری ماشین به خصوص بینایی ماشین از خود نشان داده است. بامطالعه ادبیات موضوع در فصل دوم نیز شاهد این واقعیت بودیم که در عمل بهبود عملکرد سامانه‌های تخمین جهت نگاه تنها استفاده از روش‌های استخراج ویژگی به صورت دستی و بدون کمک روش‌های یادگیری عمیق میسر نیست، کما اینکه طی چند ماه گذشته روش‌های یادگیری عمیق بهترین نتایج را در تخمین جهت نگاه به خود اختصاص داده‌اند. از طرف دیگر مدل‌های گرافیکی احتمالاتی به عنوان یک دسته‌بند قوی طی چند سال گذشته نشان داده‌اند توانایی بالایی در مدل‌سازی روابط زمانی بین فریم‌ها و روابط مکانی بین پیکسل‌های تصاویر دارند. بدین ترتیب به نظر می‌رسد ترکیب این دو روش نتایج را در تخمین جهت نگاه انسان بهبود بخشدند.

¹ Marginal probability

² maximum a posteriori probability

فصل ۴:

روش تحقیق

۱-۴ - مقدمه

در این بخش آزمایش‌های مختلفی را بیان می‌کنیم که در راستای تخمین جهت نگاه در یک سلسله مراتب زمانی به منظور بهبود دقت انجام شده است. به همین جهت ابتدا جزئیاتی درباره پیاده‌سازی آزمایش‌ها بیان می‌شود. در انتهای این بخش سامانه نهایی ارائه شده معرفی خواهد شد.

۲-۴ - جزئیات پیاده‌سازی

کار پیاده‌سازی آزمایش‌های یادگیری عمیق با استفاده از ابزار یادگیری عمیق Caffe [۴۷] انجام شده است. یک ابزار یادگیری عمیق است که توسط گروه یادگیری و بینایی ماشین دانشگاه برکلی توسعه داده شده است. برای انجام آموزش در محیط Caffe لازم است دو فایل ایجاد شوند، یک فایل معماری شبکه را نشان می‌دهد و فایل دیگر که solver نام دارد جزئیات پیاده‌سازی عملیات آموزش و اعتبارسنجی را شامل می‌شود.

در حالت کلی برای انجام آزمایش‌های یادگیری عمیق می‌توان از دو راهبرد استفاده کرد:

- یادگیری از پایه^۱
- یادگیری از طریق تنظیم وزن‌های از پیش آموزش داده شده^۲

همان‌طور که گفته شد در Caffe با استفاده از solver ها، تنظیمات مختلف مربوط اجرای آموزش و آزمایش شبکه مشخص می‌شوند. تنظیماتی نظری روش بهینه‌سازی، نرخ یادگیری^۳، تعداد تکرار آزمایش‌ها در solver مشخص می‌شوند. وظایف یادگیری بین solver و شبکه تقسیم شده است به این صورت که کارهایی نظری نظارت بر بهینه‌سازی و به روزرسانی پارامترها مربوط به solver بوده و محاسبه خطای مقادیر گرادیان به عهده شبکه (مدل شبکه) است.

روش‌های بهینه‌سازی که قابل استفاده در Solver هستند را در زیر مشاهده می‌کنید:

- Stochastic Gradient Descent (type: "SGD")
- AdaDelta (type: "AdaDelta")
- Adaptive Gradient (type: "AdaGrad")

¹ Learning from Scratch

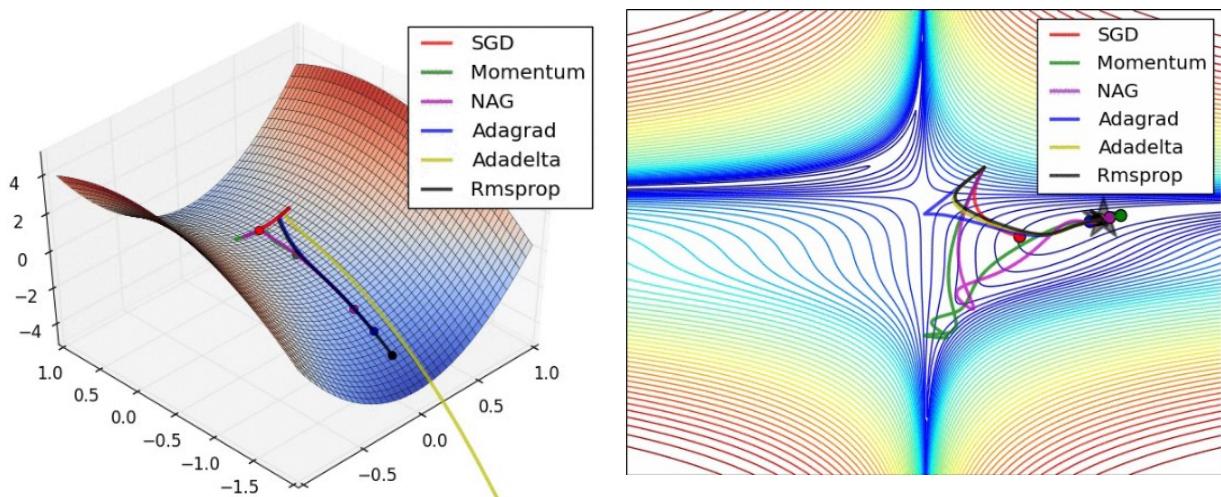
² Learning using finetune

³ Learning rate

- Adam (type: "Adam")
- Nesterov's Accelerated Gradient (type: "Nesterov")
- RMSprop (type: "RMSProp")

به طور کلی وظایف یک solver را می‌توان در موارد زیر خلاصه کرد:

۱. ساماندهی بهینه‌سازی و ایجاد شبکه آموزشی جهت یادگیری و همین‌طور ایجاد شبکه (های) آزمایشی جهت ارزیابی خروجی شبکه.
 ۲. انجام بهینه‌سازی با فراخوانی‌های پی‌درپی فازهای روبه‌جلو^۱ و عقب‌گرد^۲ و بهروزرسانی پارامترهای شبکه.
 ۳. ارزیابی (دوره‌ای) شبکه.
 ۴. گرفتن تصویر لحظه‌ای^۳ از مدل شبکه و وضعیت solver در طی بهینه‌سازی.
- در هر تکرار از عمل بهینه‌سازی، مراحل زیر اتفاق می‌افتد:
۱. فاز روبه‌جلو شبکه برای محاسبه خروجی و خطای شبکه فراخوانی می‌شود.
 ۲. فاز عقب‌گرد شبکه برای محاسبه گرادیان‌ها فراخوانی می‌شود.
 ۳. گرادیان‌ها در بهروزرسانی پارامترها با توجه به روش مشخص شده در solver اعمال می‌شوند.
 ۴. وضعیت solver با توجه به نرخ یادگیری، تاریخچه و روش مشخص شده بهروزرسانی می‌شود تا وزن‌ها را از مقداردهی اولیه به مدل یاد گرفته شده منتقل کند.



شکل (۴-۱) مقایسه سرعت همگرایی روش‌های بهینه‌سازی مختلف [۴۸]

¹ forward

² backward

³ Snapshot

۲-۳-۴ - روش‌های بهینه‌سازی

روش‌های بهینه‌سازی که در Solver مشخص می‌شوند به مسائل بهینه‌سازی کلی کاهش خطا می‌پردازند. به عنوان مثال اگر فرض کنیم ما دارای مجموعه دادگان D باشیم، هدف بهینه‌سازی میانگین‌گیری خطا از تمام $|D|$ نمونه داده در این مجموعه دادگان است

$$L(W) = \frac{1}{|D|} \sum_i^{|D|} f_w(x^{(i)}) + \lambda r(W) \quad (1-4)$$

عبارتی که در بالا مشاهده می‌کنید همان عبارت محاسبه میانگین خطای $f_w(x^{(i)})$ خطای نمونه داده $x^{(i)}$ بوده و $r(W)$ عبارت تنظیم‌کننده^۱ است که دارای وزن λ است. از آنجایی که $|D|$ می‌تواند خیلی بزرگ باشد، در عمل، در هر تکرار solver، از تقریب تصادفی^۲ این خطای استفاده می‌شود. برای این کار به جای استفاده از تمام نمونه‌های مجموعه دادگان ($|D|$) از یک دسته کوچک^۳ که تعداد نمونه‌هایی آن به مراتب بسیار کمتر از تعداد نمونه‌های موجود در مجموعه دادگان D است را به صورت زیر استفاده می‌شود:

$$L(W) \approx \frac{1}{N} \sum_i^{|D|} f_w(x^{(i)}) + \lambda r(W) \quad (2-4)$$

در اینجا مدل، f_w را در حرکت روبرو به جلو محاسبه کرده و گرادیان ∇f_w را نیز در حرکت رو به عقب محاسبه می‌کند. مقدار ΔW (به روزرسانی وزن) به وسیله solver از طریق گرادیان خطای ∇f_w ، گرادیان $\nabla r(w)$ و سایر پارامترهای اختصاصی مربوط به روش مشخص شده در آن به دست می‌آید.

SGD^۴

این روش که با نوع SGD در solver قابل استفاده است (type: "SGD") ماتریس وزن W را توسط ترکیب خطی منفی گرادیان $\nabla L(W)$ و مقدار تغییرات وزن قبلی V_t به روزرسانی می‌کند. مقدار نرخ یادگیری α برابر با وزن منفی گرادیان و مقدار μ (momentum) برابر با وزن به روزرسانی قبلی است.

¹ Regularization term

² stochastic approximation

³ mini-batch

⁴ Stochastic gradient descent

به‌طور کلی ما برای محاسبه مقدار جدید V_{t+1} و وزن‌های به‌روزرسانی شده ماتریس W_{t+1} در تکرار $t+1$ با داشتن مقدار به‌روزرسانی قبلی V_t و ماتریس وزن فعلی W_t از عبارت زیر استفاده می‌کنیم:

$$V_{t+1} = \mu V_t - \alpha \nabla L(W_t) \quad (3-4)$$

$$W_{t+1} = W_t + V_{t+1} \quad (4-4)$$

احتمالاً برای به دست آوردن بهترین نتایج، فرا پارامترهای α و μ نیازمند کمی تغییر هستند.

قواعد کلی درباره تنظیم نرخ یادگیری (α) و (μ)

یک روش برای گرفتن نتایج بهتر با استفاده از روش SGD مقداردهی نرخ یادگیری α حول مقدار 0.01 است و در صورتی که خطا شروع به افزایش کرد، کاهش مقدار آن در فاز آموزش با یک ضریب ثابت مثل 10 و تکرار این روش برای چندین مرتبه است. به‌طور کلی، توصیه می‌شود از momentum با مقدار $\mu = 0.9$ یا مقداری مشابه آن استفاده کنیم. با (هموارتر شدن) بهتر شدن به‌روزرسانی وزن‌ها در طی تکرارهای مختلف، momentum باعث سریع‌تر و پایدارتر شدن یادگیری عمیق با SGD می‌شود. [۴۳]

سیاست‌های نرخ یادگیری

در Caffe می‌توان سیاست‌های مختلفی برای کاهش نرخ یادگیری لحاظ کرد. در این قسمت فهرستی از این سیاست‌ها و نحوه عملکرد آن‌ها را مشاهده می‌کنید:

- : همیشه از base_lr استفاده می‌کند. (مقدار نرخ یادگیری ثابت است)
- : نرخ یادگیری از رابطه base_lr * gamma ^ (floor(iter / step)) به دست می‌آید.
- : نرخ یادگیری از رابطه base_lr * gamma ^ iter به دست می‌آید.
- : نرخ یادگیری از رابطه base_lr * (1 + gamma * iter) ^ (-power) به دست می‌آید.
- : همانند روش step عمل کرده با این تفاوت که اجازه تعریف گام‌های غیر یکسان در Multistep را می‌دهد.
- : در این روش، نرخ یادگیری از یک کاهش چندجمله‌ای تبعیت کرده و با رسیدن به صفر می‌شود. نرخ یادگیری در این روش از این رابطه به دست می‌آید:

$$\text{base_lr} * (1 - \text{iter}/\text{max_iter}) ^ (\text{power})$$
- : در این روش، نرخ یادگیری از یک کاهش سیگموئیدی تبعیت کرده و از Sigmoid رابطه به دست می‌آید.

$$\text{base_lr} * (1/(1 + \exp(-\gamma * (\text{iter} - \text{stepsize}))))$$

روش AdaDelta

روش AdaDelta که با نوع AdaDelta در solver قابل استفاده است (type: “AdaDelta”). یک روش بهینه‌سازی با نرخ یادگیری قدرتمند است که در سال ۲۰۱۲ ارائه شد^[۴۹]. این روش هم همانند SGD یک روش بهینه‌سازی مبتنی بر گرادیان بوده و فرمول بهروزرسانی آن بصورت زیر است:

$$(v_t)_i = \frac{RMS((v_{t-1})_i)}{RMS(\nabla L(W_t))_i} (\nabla L(W_t))_i \quad (5-4)$$

$$RMS(\nabla L(W_t))_i = \sqrt{E[g^2] + \epsilon} \quad (6-4)$$

$$E[g^2]_t = \delta E[g^2]_{t-1} + (1 - \delta) g_t^2 \quad (7-4)$$

$$(W_{t+1})_i = (W_t)_i - \alpha (v_t)_i \quad (8-4)$$

AdaGrad

روش گرادیان قابل تطبیق^۱ که با نوع AdaGrad در solver قابل استفاده است (type: “AdaGrad”). یک روش بهینه‌سازی مبتنی بر گرادیان است^[۵۰]. به تعبیر نویسنده‌گان این مقاله، این روش سعی می‌کند سوزن را در انبار کاه پیدا کند و این کار را با استفاده از ویژگی‌های بسیار قابل پیش‌بینی اما بهندرت دیده شده انجام می‌دهد^[۵۱].

با داشتن اطلاعات بهروزرسانی از تمام تکرار‌های قبلی $\nabla L(W)_{t'} \in \{1, 2, \dots, t\}$. فرمولی که توسط Duchi برای هر مولفه i از ماتریس وزن W ارائه شده است بصورت زیر می‌باشد:

$$(W_{t+1})_i = (W_t)_i - \frac{\alpha (\nabla L(W_t))_i}{\sqrt{\sum_{t'=1}^t (\nabla L(W_{t'}))_i^2}} \quad (9-4)$$

در عمل، برای وزن‌های $W \in \mathbb{R}^d$ ، روش AdaGrad به صورتی پیاده سازی می‌شود تا بجای اینکه از $O(dt)$ فضای اضافی که برای ذخیره تک تک گرادیان‌های قبلی لازم است استفاده کند تنها به اندازه $O(d)$ فضای اضافی برای اطلاعات مربوط به گرادیان‌های قبلی استفاده کند.

¹ adaptive gradient

Adam

روش Adam که با نوع solver قابل استفاده است (type: "Adam") یک روش بهینه سازی مبتنی بر گرادیان است که توسط D.Kingma ارائه شد [۵۲]. این روش را میتوان تعمیمی از روش AdaGrad دانست که فرمول محاسبه آن در زیر آمده است:

$$(m_t)_i = \beta_1(m_{t-1})_i + (1 - \beta_1)(\nabla L(W_t))_i \quad (1+4)$$

$$(v_t)_i = \beta_2(v_{t-1})_i + (1 - \beta_2)(\nabla L(W_t))_i^2 \quad (11-4)$$

$$(W_{t+1})_i = (W_t)_i - \alpha \frac{\sqrt{1 - (\beta_2)_i^t}}{1 - (\beta_1)_i^t} \frac{(m_t)_i}{\sqrt{(v_t)_i} + \varepsilon} \quad (12-4)$$

Kingma [۵۲] پیشنهاد کرده است برای مقادیر β_1 و β_2 به ترتیب از مقادیر ۰،۰،۹۹۹ و ۰،۹۹۹،۰ به ترتیب از مقادیر momentum2 و momentum Caffe استفاده شود. در این مقاله β_1 و β_2 به ترتیب معادل ϵ و $\beta_2 \cdot \epsilon$ هستند.

RMSPROP

RMSprop روشی است که با نام "RMSProp" در Solver قابل استفاده است. این روش بهینه‌سازی مبتنی بر گرادیان همانند SGD است که توسط Tieleman ارائه شده است [۵۳]. فرمول محاسبه به روزرسانی پارامترها را در زیر مشاهده می‌کنید:

$$(v_t)_i = \begin{cases} (v_{t-1})_i + \delta, & (\nabla L(W_t))_i (\nabla L(W_{t-1}))_i > 0 \\ (v_{t-1})_i \cdot (1 - \delta), & \text{else} \end{cases} \quad (13-4)$$

$$(W_{t+1})_i = (W_t)_i - \alpha(v_t)_i, \quad (14-4)$$

درصورتی که نتایج به روزرسانی‌ها دارای نوسان باشد، گرادیان با ضریب δ کاهش پیدا میکند. در غیر اینصورت به اندازه δ افزایش پیدا میکند. مقدار بیش فرض δ (rms decay) نیز بر اثر با $0.02 = \delta$ است.

۴-۳-۲- آزمایش‌ها

در این بخش آزمایش‌های مختلفی را که با استفاده از یادگیری عمیق انجام داده‌ایم با ذکر جزئیات بیان می‌کنیم.

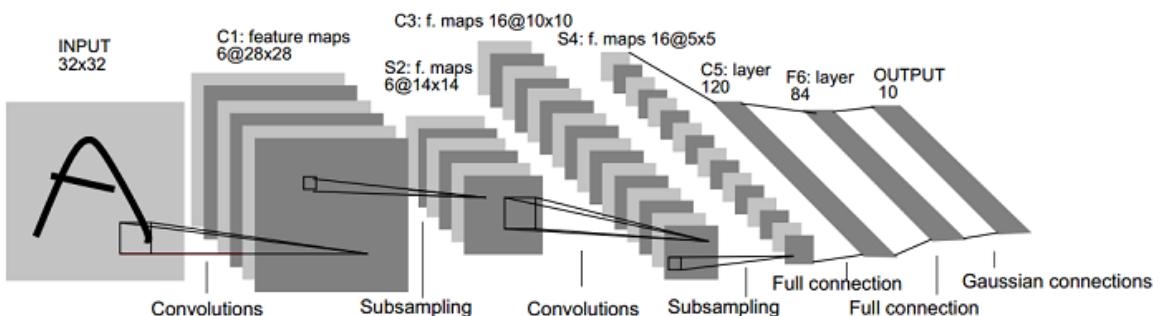
آزمایش شماره ۱ : تشخیص جهت چشم با استفاده از اطلاعات تصویر صورت و جهت سر

اولین آزمایش استفاده از اطلاعات چشم و جهت سر برای تخمین جهت نگاه و مبنای آن مقاله [۲۹] بوده است، با این تفاوت که در این آزمایش از اطلاعات هر دو چشم استفاده می‌کنیم. در این آزمایش راهبرد اولیه یادگیری عمیق مورد استفاده واقع شده است، بدین معنا که وزن‌های اولیه به صورت تصادفی مقداردهی شده‌اند و کار آموزش با این وزن‌ها آغاز می‌شود.

برای انجام این آزمایش مجموعه دادگاه EYEDIAP به ۵ دسته تقسیم شده که از این میان ۳ دسته برای آموزش، یک دسته برای اعتبارسنجی و دسته آخر برای فرآیند آزمون استفاده شده‌اند.

در این مقاله از شبکه LeNet[54] به منظور استخراج ویژگی‌های عمیق از چشم استفاده شده است. این شبکه یکی از اولین شبکه‌های عصبی پیچشی است که اولین بار در سال ۱۹۹۵[۵۵] از آن برای تشخیص اعداد دستنویس (MNIST) استفاده شده است.

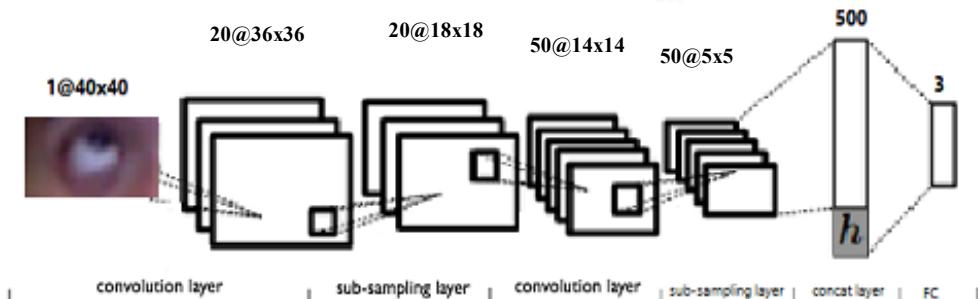
LeNet خانواده‌ای از شبکه‌های عصبی پیچشی شامل LeNet-1، LeNet-4 و LeNet-5 است. اولین نسخه از LeNet از ReLU به جای Tanh استفاده می‌کرده است.



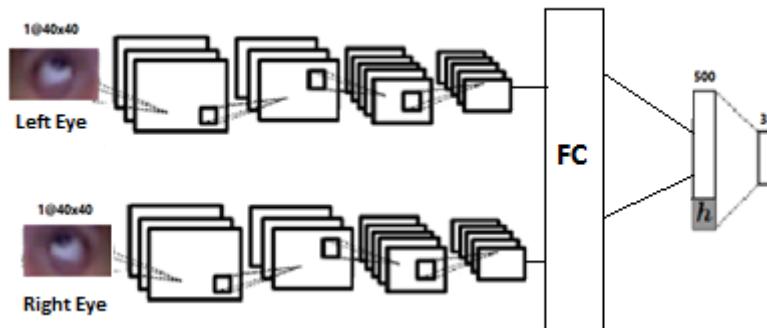
شکل (۲-۴) شبکه LeNet-5

همان‌طور که در شکل (۲-۴) مشاهده می‌کنید معماری شبکه LeNet-5 شامل یک لایه پیچش، یک

لایه ادغام ماکریم، یک لایه پیچش و یک لایه دیگر ادغام ماکریم و درنهایت لایه تماماً متصل است. شبکه مورداستفاده در این آزمایش نیز همان LeNet-5 است. این شبکه را با ورودی تصویر چشم در شکل (۳-۴) مشاهده می‌کنید:



شکل (۳-۴) شبکه LeNet تغییریافته در آزمایش شماره ۱



شکل (۴-۴) استفاده از اطلاعات هر دو چشم در آزمایش شماره یک

همان‌طور که مشاهده می‌شود ورودی این شبکه تصویر چشم در ابعاد ۴۰ در ۴۰ پیکسل است. این تصویر ابتدا از اولین لایه پیچش با اندازه فیلتر (کرنل) ۵ و اندازه لغزش^۱ ۱ پیکسل عبور می‌کند. خروجی این لایه نقشه ویژگی به اندازه ۲۰ و فیلتر به کاررفته در این لایه از نوع گاوسی و با انحراف معیار ۱،۰،۰ است. لایه دوم یک لایه ادغام از نوع ماکریم است. اندازه فیلتر برای این لایه ادغام ۲ و با لغزش ۲ پیکسل است. لایه سوم باز هم یک لایه پیچش دیگر مشابه با لایه اول است با این تفاوت که اندازه نقشه ویژگی استفاده شده برای

¹ stride

این لایه ۵۰ است. لایه چهارم نیز یک لایه ادغام ماکریم کاملاً مشابه با لایه دوم است. لایه پنجم یک لایه تماماً متصل^۱ با فیلتر نوع خاويیر^۲ [56] و اندازه نقشه ویژگی ۵۰۰ است. مقداردهی اولیه خاويیر به سیگنال کمک می‌کند در لایه‌های عمیق شبکه ادامه باید:

- اگر وزن‌ها در یک شبکه با مقادیر بسیار کوچکی آغاز شوند، سیگنال موردنظر با عبور از هر لایه ضعیفتر شده و بهندرت کوچک‌تر از آن می‌شود که بتوان از آن استفاده کرد.
 - اگر وزن‌ها در یک شبکه با مقادیر بسیار بزرگی آغاز شوند، سیگنال موردنظر با عبور از هر لایه رشد کرده و بهندرت آن قدر بزرگ می‌شود که نمی‌توان از آن استفاده کرد.
- مقداردهی اولیه با استفاده از خاويیر این اطمینان را ایجاد می‌کند که وزن‌ها در بازه صحیحی هستند و سیگنال موردنظر به لحاظ اندازه در عبور از لایه‌های مختلف معقول می‌ماند. [51]
- لایه ششم یک ReLU است. در لایه هفتم اطلاعات جهت سر از طریق یک لایه Concat با اطلاعات استخراج شده از چشم ترکیب می‌شوند. در نهایت لایه انتهایی که یک لایه تماماً متصل با اندازه خروجی ۳ است، مختصات جهت چشم را در فضای سه بعدی به دست می‌دهد.

نظر به توضیحات ارائه شده در بخش قبل، کار آموزش ابتدا با استفاده از روش بهینه‌سازی SGD انجام شده است. بدین ترتیب متغیرهای زیر در Solver با مقادیر ذکر شده مقداردهی شده‌اند:

```
base_lr: 0.01
lr_policy: "step"
gamma: 0.1
stepsize: 100000
max_iter: 350000
momentum: 0.9
```

در خط اول مشخص می‌کنیم که آموزش با نرخ یادگیری 10^{-4} شروع شود. در خط دوم سیاست نرخ یادگیری را مشخص کرده‌ایم، در این قسمت ما مشخص کردیم که نرخ یادگیری طی گام‌هایی کاهش یابد. در ادامه بیشتر در این مورد توضیح داده شده است. خط سوم، همان ضریب کاهش نرخ یادگیری است که در هر گام انجام می‌شود. در این قسمت مشخص می‌کنیم نرخ یادگیری با چه ضریبی کاهش پیدا کند. در اینجا ما مشخص کردیم که نرخ یادگیری با ضریب 10^{-4} کاهش پیدا کند (عنی نرخ یادگیری در مقدار gamma که مساوی 10^{-4} است ضرب شود).

خط چهارم همان تعداد گام‌هایی است که نرخ یادگیری در آن‌ها باید کاهش یابد. در اینجا این مقدار برابر با ۱۰۰ هزار تکرار، نرخ یادگیری کاهش پیدا می‌کند. خط پنجم تعداد تکرار

¹ Fully Connected

² Xavier

مراحل آموزش را مشخص می‌کند. در اینجا ۳۵۰ هزار مرتبه عملیات آموزش تکرار شده است. خط آخر نیز مقدار momentum را مشخص کرده است.

در این آزمایش، کار آموزش را با نرخ یادگیری پایه $\alpha = 0.01$ برای ۱۰۰ هزار تکرار اول شروع کرده‌ایم و سپس نرخ یادگیری را با مقدار γ ضرب کرده و آموزش را با نرخ یادگیری 0.001 برای ۱۰۰ هزار تکرار بعدی ($200 - 100$) ادامه داده‌ایم. به همین صورت برای تکرارهای ۲۰۰ هزار تا ۳۰۰ هزار از نرخ یادگیری 1×10^{-4} استفاده کرده و نهایتاً از نرخ یادگیری 1×10^{-6} برای تکرارهای باقی‌مانده ($300 - 250$ هزار) استفاده کرده‌ایم. نتیجه به دست‌آمده شامل واگرایی و یا دقت ناکافی بر روی بخش اعتبارسنجی بود. با تغییر پارامترها نیز همگرایی مورد نظر حاصل نشده است.

بدین ترتیب کار آموزش بار دیگر با استفاده از روش Adadelta که سرعت همگرایی به مراتب بهتری نسبت به SGD دارد بر روی ۱۶۰ هزار فریم، با نرخ یادگیری 1×10^{-4} و به تعداد ۸۰ هزار تکرار انجام شد. پارامتر δ : $1e-6$ در این روش استفاده شده و همگرایی مناسبی بر روی داده‌های اعتبارسنجی حاصل شد.

آزمایش شماره ۲: تشخیص جهت چشم با استفاده از اطلاعات استخراج شده به روش عمیق از روی جهت سر

اشکال آزمایش قبل استفاده از اطلاعات جهت سر به صورت خام است که در مجموعه دادگان موجود بودند. در شرایط واقعی این اطلاعات در دسترس نیستند. همین امر موجب طراحی این آزمایش شده است تا اطلاعات سر به طور خودکار و با استفاده از یادگیری عمیق قابل دسترسی باشند.

شبکه استفاده شده در این آزمایش همان شبکه LeNet مورد استفاده در آزمایش قبل است با این تفاوت که برای انجام این آزمایش از فریم‌های خام به عنوان ورودی استفاده شده است. به علاوه لایه concat نیز در این آزمایش حذف شده است. برچسب داده‌ها در اینجا یک بردار ۱۲ تایی است شامل ماتریس چرخش و ماتریس انتقال که هر کدام به ترتیب ۹ و ۳ درایه دارند.

تعریف: ماتریس چرخش برای چرخش به اندازه α رادیان حول محور x ، β رادیان حول محور y و γ رادیان حول محور z به صورت زیر تعریف می‌شود:

$$R_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}$$

¹ iteration

$$R_y(\beta) = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix}$$

$$R_z(\gamma) = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

بنابراین اگر عملیات چرخش را ابتدا حول محور x ، سپس حول محور y و درنهایت حول محور z انجام دهیم، این مجموعه چرخش را می‌توان به صورت حاصل ضرب $R = R_x(\alpha) R_y(\beta) R_z(\gamma)$ در نظر گرفت که معادل است با ماتریس چرخش زیر :

$$\begin{bmatrix} \cos\beta\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \cos\beta\sin\gamma & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma \\ -\sin\beta & \sin\alpha\cos\beta & \cos\alpha\cos\beta \end{bmatrix}$$

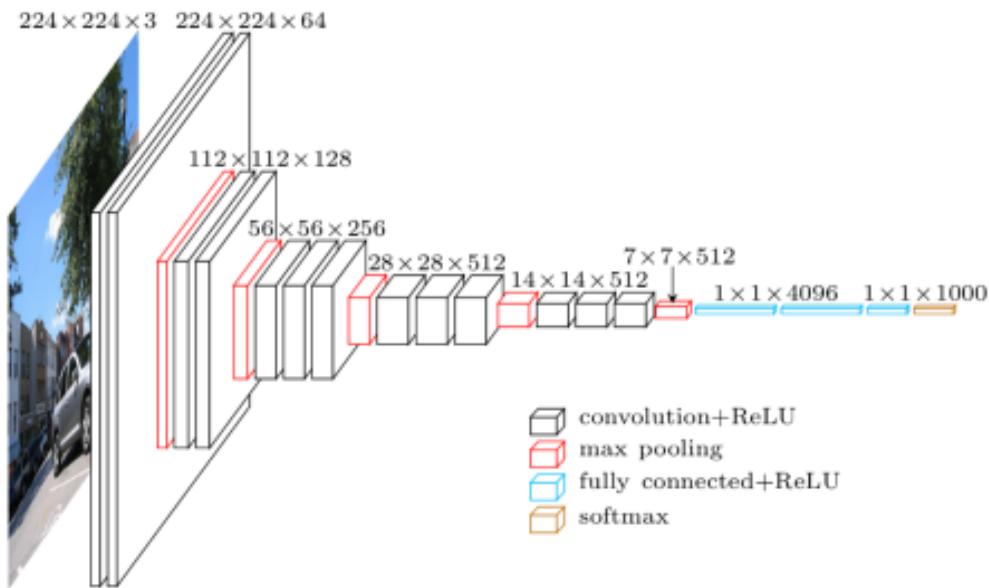
برای انجام این آزمایش مجموعه دادگان به دو بخش مساوی تقسیم شده است. فرآیند آموزش بر روی بخش اول انجام شده و اطلاعات سر شامل بردارهای ۱۲ تایی برای بخش دوم مجموعه دادگان تخمین زده می‌شود.

سپس کار تشخیص جهت نگاه مجدداً و با این اطلاعات به دست آمده انجام می‌شود. بدین ترتیب برای آموزش، اعتبارسنجی و آزمون در این مرحله نیمی از مجموعه دادگان در دسترس است. این تعداد از داده‌ها را به ۶ بخش تقسیم کرده، ۴ بخش را به آموزش اختصاص داده و دو بخش دیگر را هر کدام یکی برای اعتبارسنجی و آزمون استفاده کرده‌ایم.

آزمایش شماره ۳: تشخیص جهت نگاه با استفاده از اطلاعات جهت سر-بهبود دقت
 هدف از طراحی این آزمایش بهبود دقت در جهت کاهش میانگین خطای تخمین جهت نگاه بوده است. بدین منظور شبکه استفاده شده در آزمایش قبل را اندکی تغییر داده‌ایم، به طوری که پیش از آخرین لایه پیچش یک لایه پیچش دیگر را با اندازه فیلتر ۲ و لغزش ۲ اضافه کرده‌ایم که اندازه خروجی ۱ دارد. بدین ترتیب ورودی این لایه ۵۰ عدد نقشه ویژگی با ابعاد $14*14$ هستند که به ۱ نقشه ویژگی با ابعاد $10*10$ تبدیل می‌شوند. با تغییر ایجاد شده خروجی دومین لایه ادغام ماکریم یک نقشه ویژگی به ابعاد $5*5$ خواهد داشت. لایه‌های تماماً متصل مانند حالات قبل ۵۰۰ و ۳ عدد را به عنوان خروجی برمی‌گردانند.

آزمایش شماره ۴: تاثیر افزایش تعداد لایه‌ها بر دقت خروجی

در این بخش هدف بررسی تاثیر افزایش تعداد لایه‌ها بر دقت تخمین جهت نگاه است. بدین‌منظور آزمایش اول مطرح شده در این بخش را این‌بار با استفاده از شبکه عمیق VGG-16 تکرار کرده‌ایم. این شبکه در سال ۲۰۱۴ توسط محققان دانشگاه آکسفورد ارائه شد.^[۵۷] معماری این شبکه را در شکل (۴-۴) مشاهده می‌کنید.



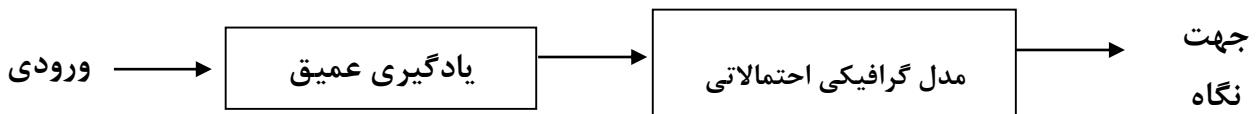
شکل (۴-۴) معماری شبکه VGG-16 [۵۸]

در این آزمایش همانند آزمایش اول ورودی شبکه تصاویر چشم به ابعاد 40×40 پیکسل هستند. علاوه بر این تصاویر اطلاعات استخراج شده از مجموعه دادگان پیرامون جهت سر نیز پس از آخرین لایه پیچش به این شبکه وارد می‌شود. تغییر دیگری که در شبکه VGG-16 داده‌ایم تعداد نورون خروجی است. همان‌طور که در شکل (۴-۴) مشاهده می‌کنید تعداد این خروجی‌ها در VGG-16، ۱۰۰۰ عدد است که در این آزمایش این تعداد به ۳ تغییر پیدا کرده است. تعداد ۱۰۰۰ خروجی در شبکه VGG به دلیل وجود ۱۰۰۰ طبقه مختلف بوده است. از آن‌جا که در این‌جا هدف تشخیص مختصات سه‌بعدی است تعداد خروجی ۳ درنظر گرفته شده است.

۴-۳-۴ - معماری سامانه

با توجه به مطالبی که به صورت مجزا در بخش‌های مختلف به آن‌ها اشاره شد، در این بخش معماری کلی

سامانه در قالب یک نگاه کلی بیان شده است. سامانه تشخیص جهت نگاه پیشنهادی در این پایان‌نامه بر اساس مفاهیم مطرح شده در حوزه‌های شبکه‌های عمیق و مدل‌های گرافیکی پایه‌ریزی شده است، به همین منظور در ادامه معماری زیر بخش‌های یادگیری عمیق و مدل گرافیکی نیز به صورت مجزا مطرح خواهند شد.



شکل (۴-۴) معماری سامانه پیشنهادی

همان‌طور که در شکل (۴-۵) نیز مشاهده می‌شود سامانه پیشنهادی از دو بخش اصلی تشکیل شده است:

- بخش یادگیری عمیق

- بخش مدل گرافیکی احتمالاتی

به طور خلاصه می‌توان گفت کار بخش اول یادگیری و استخراج ویژگی‌های موردنیاز است. این ویژگی‌ها به بخش دوم در تشخیص جهت چشم کمک می‌کنند.

۴-۳-۲- معماری بخش یادگیری عمیق

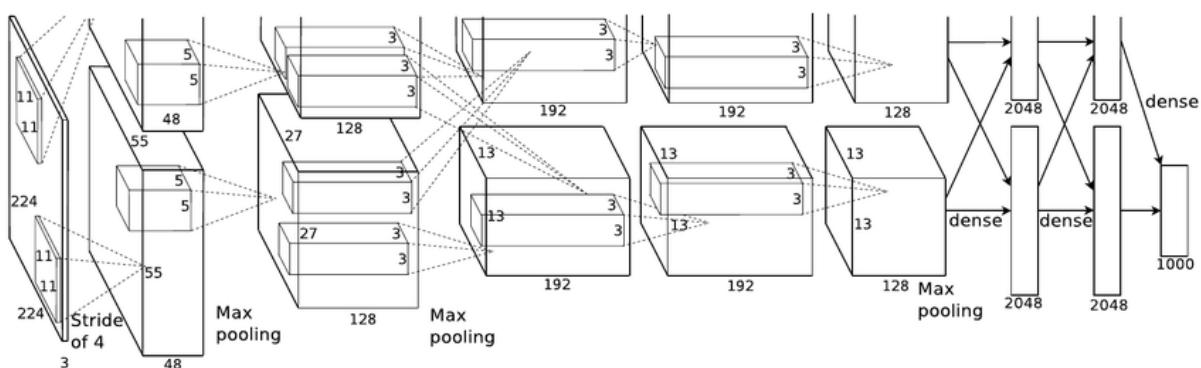
در این بخش معماری بخش ابتدایی سامانه پیشنهادی یعنی زیر بخش یادگیری عمیق مطرح شده است. همان‌طور که در قسمت قبل بیان شد وظیفه این بخش یادگیری ویژگی‌هایی است که به مدل گرافی در تشخیص جهت چشم کمک می‌کنند.

پس از انجام آزمایش‌های متعددی که با ذکر جزئیات در بخش‌های قبلی مطرح شدند و نظر به یافته‌های [۳۰، ۳۱] در این بخش نیز سعی کرده‌ایم با استفاده از یادگیری عمیق اطلاعات موجود در صورت را استخراج کنیم تا به کمک این اطلاعات تشخیص جهت نگاه انسان انجام شود.

شبکه معرفی شده در این بخش الهام گرفته از شبکه معروف AlexNet [۴۳] است که در دسته‌بندی تصاویر

استفاده می‌شود. این شبکه برای نخستین بار در سال ۲۰۱۲ و در رقابت‌های^۱ ILSVRC معرفی شد. به دلیل اهمیت این شبکه در ارائه معماری نهایی به ذکر جزئیات این شبکه می‌پردازیم.

شکل شبکه AlexNet به صورت زیر است:



شکل (۷-۴) معماری شبکه AlexNet[43]

- از آن‌جا که آموزش این شبکه بر روی ۲ عدد GPU انجام شده است شکل شبکه دارای دو بخش موازی است.
- این شبکه ۱۱ لایه دارد.
- نویسنده‌گان این مقاله شبکه ارائه شده را بر روی ۱۵ میلیون تصویر دارای برچسب که در ۲۲ هزار کلاس مختلف قرار داشتند آموزش دادند.
- آن‌ها از تابع فعال‌سازی ReLU استفاده کرده‌اند. آن‌ها هم‌چنین از لایه‌های dropout برای مقابله با مشکل آموزش بیش از حد^۲ استفاده کرده‌اند.
- نویسنده‌گان مقاله برای آموزش از بهینه‌سازی^۳ SGD استفاده کرده‌اند.
- آن‌ها برای آموزش شبکه AlexNet خود از دو عدد GPU مدل GTX 580 استفاده کرده و اعلام کردند که کار آموزش حدود ۶ روز به طول انجامیده است.
- برای نخستین بار یک مدل توانسته بود به خوبی بر روی مجموعه‌دادگان ImageNet عملیات دسته‌بندی را انجام دهد. در جریان این رقابت و معرفی شبکه AlexNet، دقت به دست آمده بر روی این مجموعه‌دادگان ۱۵,۴ درصد^۴ گزارش شد.

¹ ImageNet Large-Scale Visual Recognition Challenge

² overfitting

³ Stochastic Gradient Descend

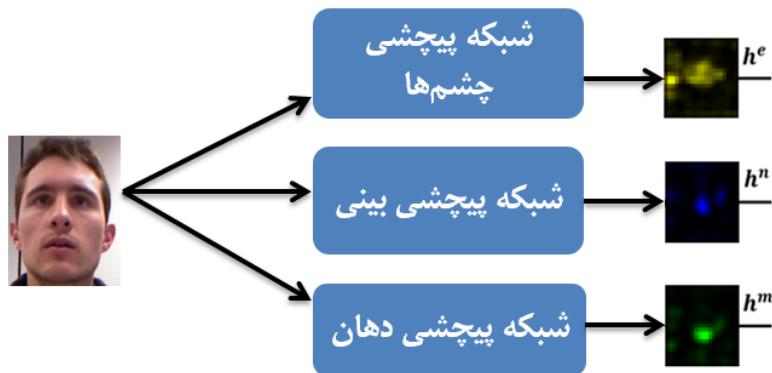
⁴ top 5 test error

جدول (۱-۴) جزئیات لایه‌های مختلف AlexNet

	۱	۲	۳	۴	۵	۶	۷	۸
نوع	conv+max+norm	conv+max+norm	conv	conv	conv+max	full	full	full
کانال‌ها	۹۶	۲۵۶	۳۸۴	۳۸۴	۲۵۶	۴۰۹۶	۴۰۹۶	۱۰۰۰
اندازه فیلتر	۱۱*۱۱	۵*۵	۳*۳	۳*۳	۳*۳	-	-	-
لغزش پیچش	۴*۴	۱*۱	۱*۱	۱*۱	۱*۱	-	-	-
اندازه ادغام	۳*۳	۳*۳	-	-	۳*۳	-	-	-
لغزش ادغام	۲*۲	۲*۲	-	-	۲*۲	-	-	-
اندازه padding	۲*۲	۱*۱	۱*۱	۱*۱	۱*۱	-	-	-

خلاصه وضعیت لایه‌های مختلف AlexNet را در جدول (۲-۳) مشاهده می‌کنید. برای مثال اندازه ویلتر پیچش در لایه اول $11*11$ پیکسل است که با اندازه لغزش 4 بر روی تصویر اصلی اعمال می‌شود. معماری کلی بخش یادگیری عمیق را در سامانه پیشنهادی در شکل (۷-۴) مشاهده می‌کنید. همان‌طور که در این شکل نیز مشاهده می‌شود شبکه یادگیری عمیق شامل سه زیر شبکه به شرح زیر است:

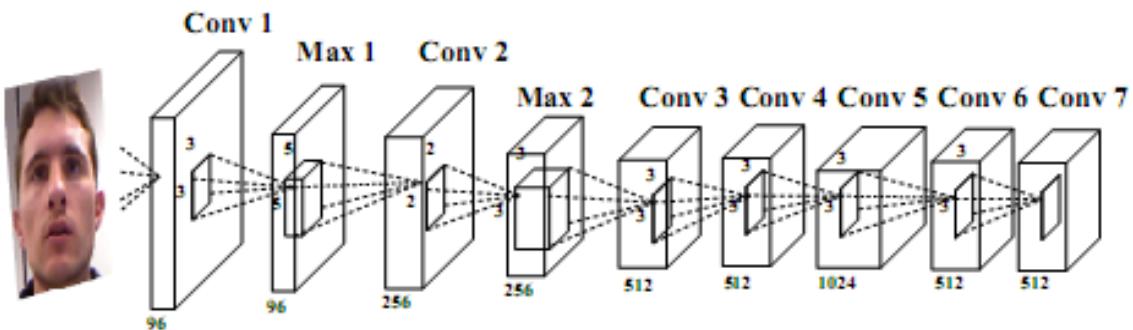
- شبکه پیچش چشم‌ها
- شبکه پیچش بینی
- شبکه پیچش دهان



شكل (۸-۴) معماری بخش یادگیری عمیق از سامانه پیشنهادی

هر کدام از این شبکه‌ها یک شبکه عصبی عمیق پیچشی^۱ است که بر مبنای AlexNet تعریف شده و بر اساس [۵۹] استفاده شده‌اند.

شبکه پیچشی هر کدام از بخش‌ها مطابق شکل زیر است:

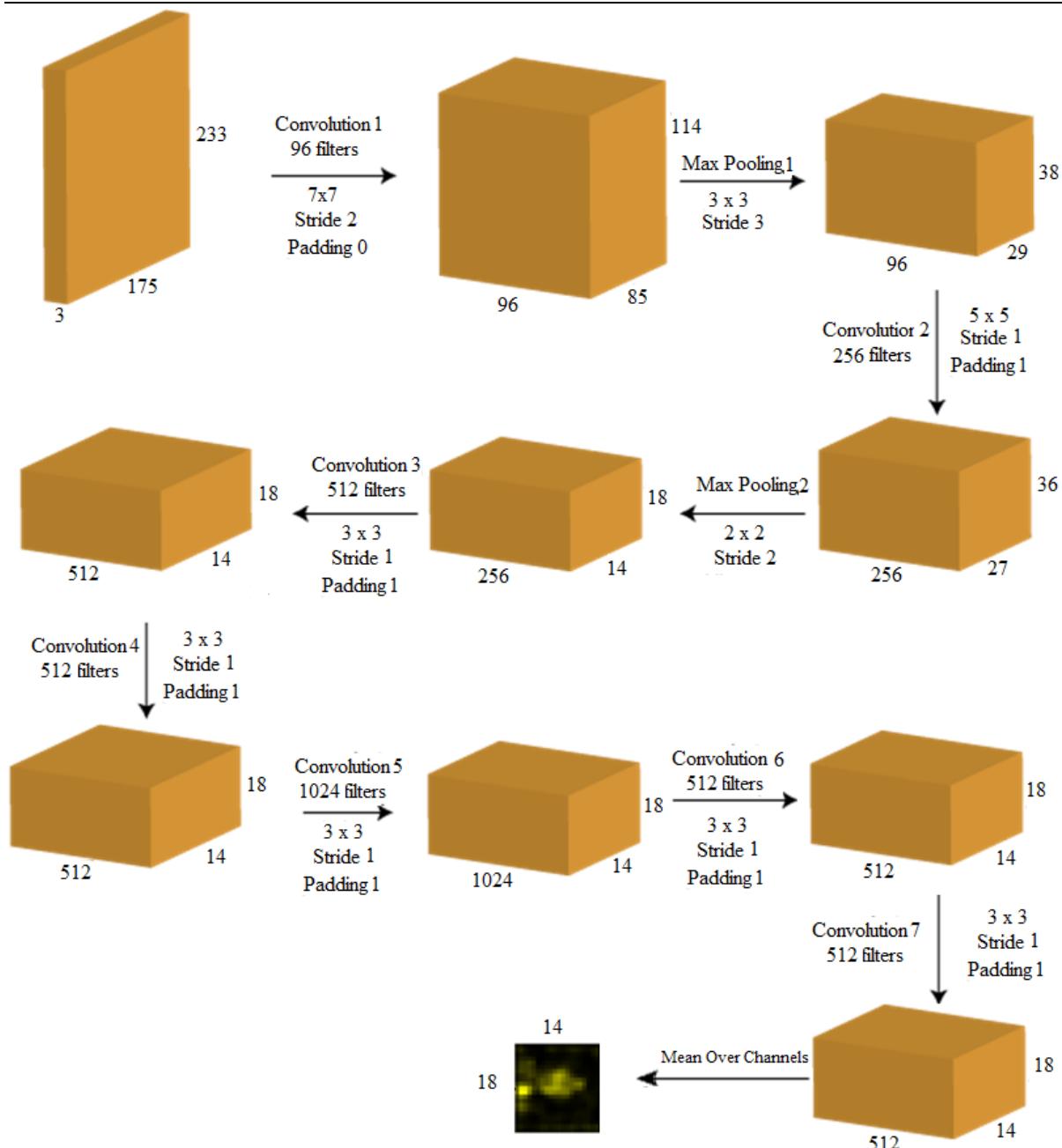


شكل (۹-۴) شبکه پیچشی استفاده در زیر بخش یادگیری عمیق

ورودی مازول یادگیری عمیق را تصویر صورت در نظر گرفته‌ایم. این تصویر از فریم خام و با کمک اطلاعات موجود در مجموعه دادگان شامل محل قرارگیری چشم‌ها استخراج شده است. همان‌طور که در شکل (۸-۴) نیز دیده می‌شود بخش‌های اصلی این شبکه ۷ لایه پیچش و دو لایه ادغام ماکریم هستند. مطالعات اخیری همچون [۶۰] نشان داده‌اند استفاده از چندین لایه پیچش مانند آنچه در شبکه AlexNet رخ می‌دهد می‌تواند به خوبی مکان اجزا را مدل‌سازی کند.

شكل (۹-۴) معماری یک شبکه عصبی پیچشی را زیر بخش یادگیری عمیق به خوبی نشان می‌دهد:

¹ Convolutional neural network

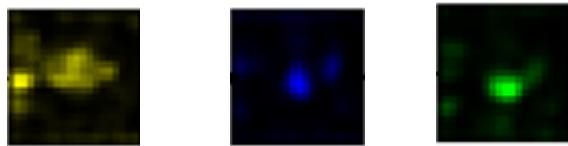


شکل (۴-۱۰) جزئیات شبکه عمیق استفاده شده در سامانه پیشنهادی

همان‌طور که در شکل (۹-۴) مشاهده می‌شود ورودی شبکه، تصویر صورت با ابعاد 233×175 پیکسل است. خروجی این تصویر با عبور از لایه‌های مختلف و درنهایت میانگین‌گیری بر روی کanal‌های مختلف نقشه ویژگی یک ماتریس 14×14 شامل ۲۵۲ درایه است.

بدین ترتیب هر تصویر صورت به شبکه‌هایی وارد می‌شود که دارای وزن‌های خاص آموزش داده شده برای استخراج چشم، بینی و دهان هستند. شکل زیر یک نمونه از تصویر ماتریس خروجی را به ازای این سه

شبکه نشان می‌دهد:



شکل (۱۱-۴) یک مثال از خروجی شبکه عمیق استفاده شده در سامانه پیشنهادی

به عبارت دقیق‌تر، خروجی [۵۹] سه عدد فایل شامل وزن‌ها از پیش آموزش داده شده است که هر کدام از این وزن‌ها برای استخراج چشم، بینی و دهان آموزش داده شده‌اند. در اینجا یک فریم تصویر صورت به عنوان ورودی به شبکه داده شده تا با استفاده از این وزن‌ها، شبکه نشان داده شده در شکل (۱۰-۴) خروجی مناسب با هر یک از این سه عضو صورت را حاصل کند.

آزمایش شماره ۵: استفاده از رگرسیون خطی به عنوان دسته‌بند

در این بخش ویژگی‌های استخراج شده از شبکه عمیق به یک رگرسیون خطی به عنوان دسته‌بند داده می‌شوند. برای پیاده‌سازی این دسته‌بند از دستور `fitrlinear` در متلب استفاده کرده‌ایم. اینتابع در متلب ۲۰۱۶ پیاده‌سازی شده و قابل استفاده است. بدین ترتیب دستور `(Y, X) = fitrlinear(X)` یک مدل رگرسیون را به عنوان پاسخ بر می‌گرداند که شامل برآنش مدل رگرسیون ماشین بردار پشتیبان برای داده‌های X و برچسب‌های Y است.

آزمایش شماره ۶: استفاده از ماشین بردار پشتیبان رگرسیون به عنوان دسته‌بند

در این بخش ویژگی‌های استخراج شده از شبکه عمیق به یک ماشین بردار پشتیبان رگرسیون به عنوان دسته‌بند داده می‌شوند. برای پیاده‌سازی این دسته‌بند از دستور `fitrsvm` در متلب استفاده کرده‌ایم.

۴-۳-۳-۴- معماری بخش مدل گرافیکی احتمالاتی

در این بخش معماری بخش دوم در سامانه پیشنهادی یعنی زیر بخش مدل گرافیکی مطرح شده است. ورودی این بخش ویژگی‌های استخراج شده از فریم ورودی است که توسط بخش یادگیری عمیق استخراج

شده است. همان‌طور که در بخش قبل توضیح داده شد این ویژگی‌ها شامل یک ماتریس 18×14 ^۱ یعنی ۲۵۲ پیکسل است.

در اینجا از مدل گرافیکی احتمالاتی با نام میدان‌های تصادفی شرطی نهان^۲ استفاده شده است. ساده‌ترین مدل میدان‌های تصادفی شرطی از مجموعه مدل‌های گرافیکی احتمالاتی بدون جهت میدان‌های تصادفی شرطی با زنجیره خطی^۳ است. آموزش یک مدل CRF به معنای پیدا کردن بردار وزن‌های w است به‌گونه‌ای که بهترین پیش‌بینی ممکن را برای هر نمونه آموزشی x ارائه دهد:

$$\bar{y}^* = \operatorname{argmax}_{\bar{y}} p(\bar{y} | \bar{x}; w) \quad (15-4)$$

نماد بالای x به معنای برداری از همه x ‌ها است.
با داشتن هر \bar{y} , x باید مقدار زیر را ارزیابی کنیم:

$$p(\bar{y} | \bar{x}; w) = \frac{1}{Z(\bar{x}, w)} e^{\sum_j w_j F_j(\bar{x}, \bar{y})} \quad (16-4)$$

که مخرج کسر همانتابع بخش‌بندی^۳ به شرح زیر است:

$$Z(\bar{x}, w) = \sum_{\bar{y}'} e^{\sum_j w_j F_j(\bar{x}, \bar{y}')} \quad (17-4)$$

در مورد میدان‌های تصادفی شرطی با زنجیره خطی هر تابع خصیصه تنها به دو برچسب که کنار هم قرار دارند وابسته است و این حقیقت فرآیند آموزش را آسان‌تر می‌کند.

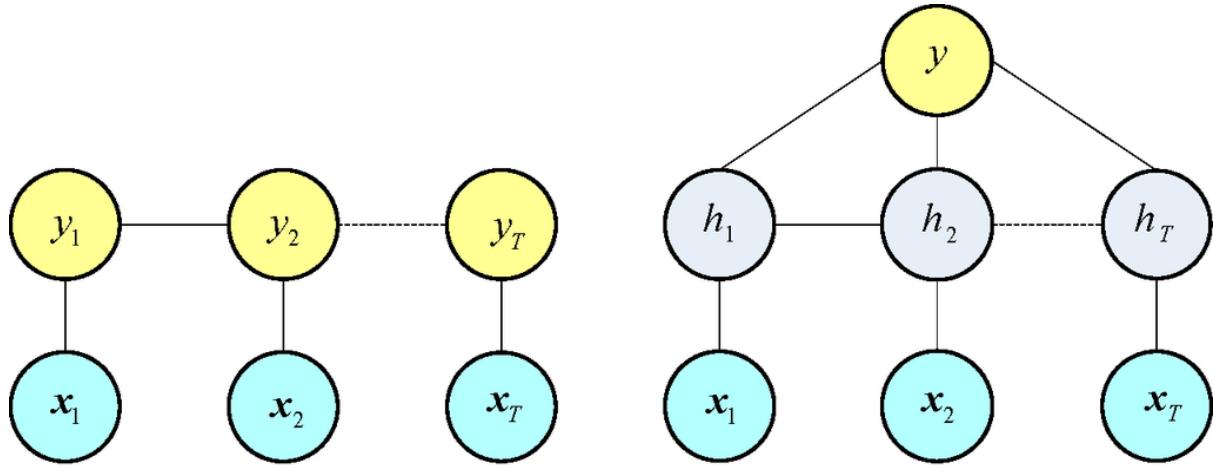
وقتی که مجموعه‌ای از نمونه‌های آموزشی را در اختیار داریم، فرض می‌کنیم هدف ما پیدا کردن پارامتر w_j است که احتمال شرطی نمونه‌های آموزشی را بیشینه می‌کند. برای این منظور نیاز است مشتق جزئی معیار شباهت شرطی را برای یک نمونه آموزشی برای هر w_j محاسبه کنیم. بیشینه کردن p همان بیشینه کردن $\ln p$ است.

¹ Hidden Conditional Random Field

² Linear-Chain CRF

³ Partition function

$$\begin{aligned} \frac{\partial}{\partial w_j} \ln p(y|x; w) &= F_j(x, y) - \frac{\partial}{\partial w_j} \log Z(x, w) \\ &= F_j(x, y) - E_{y': p(y'|x; w)} [F_j(x, y')]. \end{aligned} \quad (18-4)$$

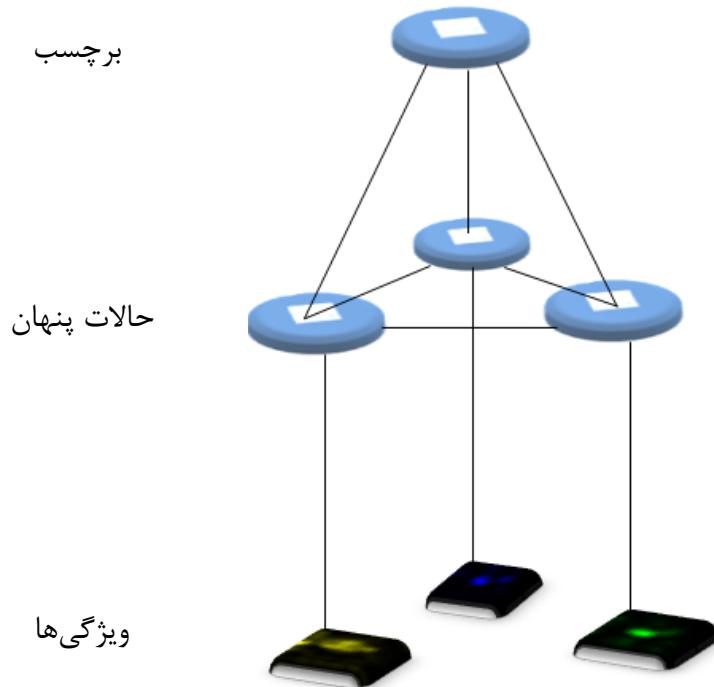


شکل (۱۲-۴) نمایی ساده از میدان‌های تصادفی شرطی (سمت چپ)-میدان‌های تصادفی شرطی نهان (سمت راست)

همان‌طور که در شکل (۱۱-۴) نیز مشاهده می‌شود مدل میدان‌های تصادفی شرطی نهان مشابه با میدان‌های تصادفی شرطی است با این تفاوت که برای هر مجموعه ویژگی x مجموعه‌ای متناهی با نام H از برچسب‌های نهان وجود دارند. مجموعه‌ای از این برچسب‌ها یک برچسب نهایی را تحت عنوان مشاهده در بخش آموزش گرفته و این برچسب را در مرحله آزمون استنتاج می‌کنند.^[۶۱] علت اصلی استفاده از حالت نهان مدل کردن تفاوت‌های درون کلاسی^۱ است.

با ذکر این توضیحات، معماری کلی مدل گرافیکی استفاده شده در این بخش به صورت زیر است:

¹ Intra-class



شکل (۱۳-۴) معماری زیر بخش مدل‌های گرافیکی احتمالاتی در سامانه پیشنهادی

همان‌طور که در شکل (۱۲-۴) نیز مشاهده می‌شود، ویژگی‌های استخراج شده از مرحله قبل به گره‌هایی متصل شده‌اند که نمایانگر حالات پنهان هستند. بدین ترتیب برای هر فریم ورودی ۳ بردار ۲۵۲ تایی ویژگی داریم. هر بردار ویژگی به گره حالت پنهان مربوط به خود متصل است. گره‌های حالات پنهان هر سه با سه یال مجزا به گره برچسب متصل شده‌اند. این برچسب همان نقطه‌ای است که بیننده به آن نگاه می‌کند. در این مدل سه نوع تابع پتانسیل^۱ قابل تعریف هستند:

- توابع پتانسیل یکه^۲
- توابع پتانسیل دوتایی^۳
- توابع پتانسیل سه‌تایی^۴

توابع پتانسیل یکه بین هر یک گره‌های ویژگی (چشم، بینی یا دهان) و گره حالت مخفی تعریف می‌شود. بنابراین برای هر فریم از مجموعه دادگان ۳ عدد تابع پتانسیل یکه داریم. توابع پتانسیل دوتایی بین هر حالت نهان و برچسب تعریف می‌شود. بعلاوه بین هر دو حالت پنهان نیز تابع پتانسیل دوتایی داریم.

¹ Potential function

² Unary

³ Pairwise

⁴ Triple

بنابراین برای هر فریم درمجموع ۶ عدد تابع پتانسیل دوتایی داریم، توابع پتانسیل سه‌تایی نیز بین برچسب و دو حالت پنهان قابل تعریف هستند. بدین ترتیب در این مدل برای هر فریم ۳ عدد تابع پتانسیل سه‌تایی وجود دارند.

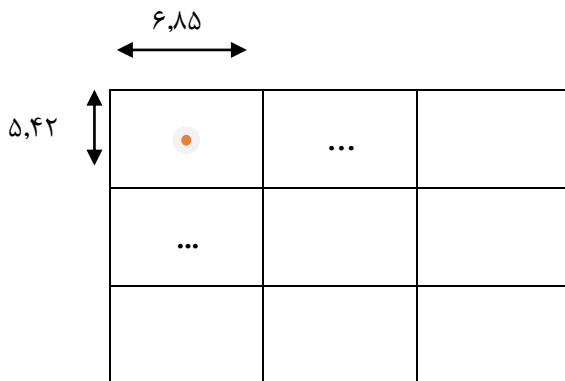
در اینجا تعداد حالات پنهان برای هر کدام از سه گره ۸ حالت در نظر گرفته شده است.

آماده‌سازی داده

از آنجاکه ذات مسئله رگرسیون و نه دسته‌بندی بوده، پس از استخراج ویژگی‌ها در مرحله قبل، باید برچسب هر فریم بر حسب کلاس تعیین شود. طبق اطلاعات موجود در مجموعه دادگان، نمایشگر مورد استفاده در تهیه داده‌ها یک نمایشگر ۲۴ اینچی بوده است. بنابراین ابعاد این نمایشگر در حدود 38×48 سانتی‌متر هستند. بدین ترتیب با تعریف ۴۹ کلاس می‌توان مستطیل‌هایی را به مساحت ۳۶ سانتی‌متر مربع بر روی صفحه نمایش در نظر گرفت. بدین ترتیب وضوح دقت در حدود ۴,۵ سانتی‌متر خواهد بود. از آنجاکه بیننده در فاصله‌ای بین ۸۰ تا ۹۰ سانتی‌متر از نمایشگر قرار دارد، وضوحی تقریباً ۳ درجه‌ای را خواهیم داشت.

$$\text{درجه} = \alpha \cdot 2 \tan^{-1}(2.5/90) = 3$$

بنابراین تعداد کلاس‌ها در این حالت ۴۹ کلاس در نظر گرفته شده و با استفاده از الگوریتم Kmeans به هر فریم یک برچسب بین ۱ تا ۴۹ اختصاص داده شده است.



شکل (۱۴-۴) بخش‌بندی نمایشگر استفاده شده در مجموعه دادگان

نکته قابل توجه در اینجا این است که با در نظر گرفتن وضوح ۳ درجه‌ای، بهترین جواب قابل انتظار از این الگوریتم ۳ درجه خطا خواهد داشت که این میزان خطا به دلیل مدل‌سازی مذکور با توجه به شکل (۱۴-۴) ایجاد شده است.

کل فریم‌های مجموعه دادگان به ۶ قسمت تقسیم شده‌اند که از ۴ بخش برای آموزش و از ۲ بخش باقی‌مانده هرکدام یک بخش برای اعتبارسنجی و آزمون استفاده شده‌اند.

با در نظر گرفتن مدل گرافیکی موردنظر مطابق با شکل (۱۲-۴) می‌توان این مدل را آموزش داد. کار آموزش این مدل با استفاده از ویژگی‌های استخراج شده از مرحله یادگیری عمیق، انجام می‌شود. مراحل آموزش در این بخش به شرح زیر هستند:

۱. ساخت مدل گرافیکی از روی داده‌ها
۲. ساخت توابع پتانسیل
۳. آموزش

کار آموزش و آزمون مدل گرافیکی ارائه شده با استفاده از ابزاری تحت عنوان^۱ dSP [۶۲] انجام می‌شود. این ابزار یک پیاده‌سازی از الگوریتم‌های یادگیری و استنتاج در مدل‌های گرافیکی است. این ابزار با کمک MPI [۶۳] مکانیسم‌های موازی‌سازی پردازش را نیز فراهم کرده است.

۴-۴- نتیجه‌گیری

در این فصل، نخست به توضیح مفاهیمی پیرامون شبکه‌های عصبی، شبکه‌های عصبی عمیق و مدل‌های گرافیکی پرداختیم. سپس آزمایش‌هایی را بیان کردیم که هرکدام از روشی خاص سعی کرده‌اند دقیق تخمین جهت نگاه را بهبود بخشنند. درنهایت معماری سامانه ارائه شده در دو زیر بخش یادگیری عمیق و مدل گرافیکی احتمالاتی معرفی و مورد بررسی قرار گرفت.

¹ Distributed Structured Prediction

فصل ۵:

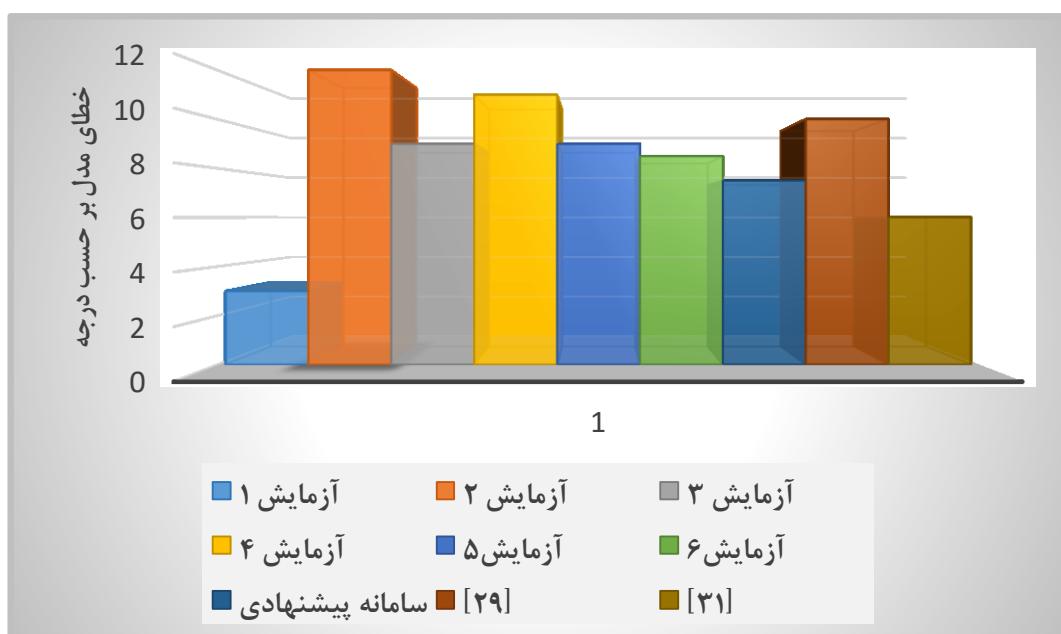
نتایج و تفسیر آن‌ها

۱-۵ - مقدمه

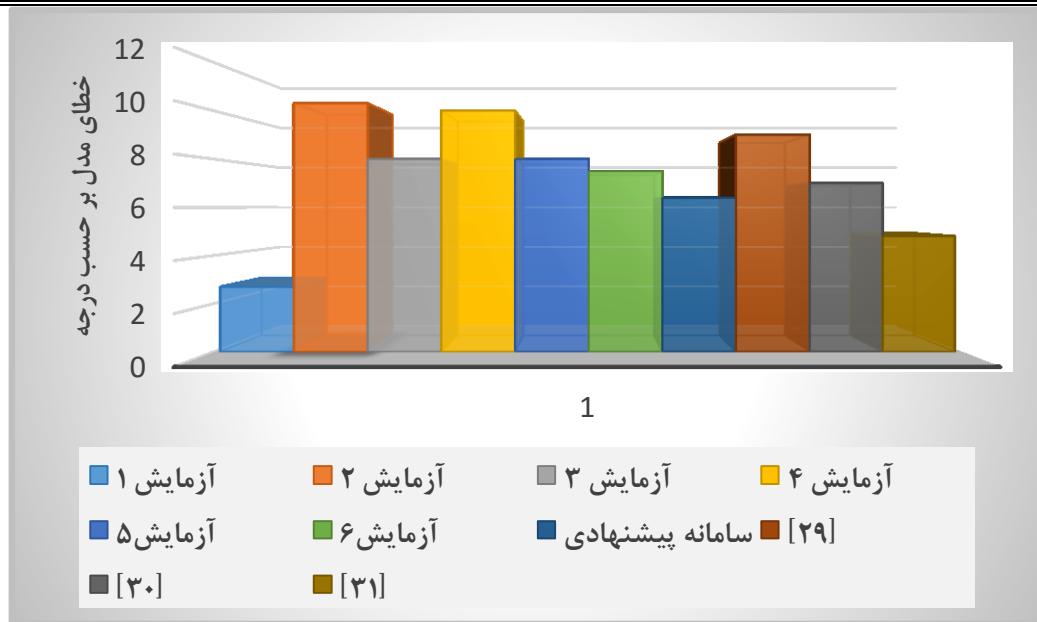
در فصل قبل علاوه بر سامانه پیشنهادی، آزمایش‌های متعددی نیز بیان شدند که سعی شد جزئیات هر کدام در طراحی آزمایش موردنظر بیان شوند. در این بخش به بررسی نتایج به دست آمده اعم از نتایج حاصل از آزمایش‌های مختلف و همچنین نتایج به دست آمده از ارائه سامانه نهایی می‌پردازیم.

۲-۵ - نتایج

در شکل‌های (۲-۵) و (۱-۵) از سمت چپ به ترتیب آزمایش‌های ۱ تا ۶، سامانه پیشنهادی و دقت‌های به دست آمده از [۲۹]، [۳۰] و [۳۱] نشان داده شده‌اند.



شکل (۱-۵) مقایسه دقت روش‌های مختلف بر روی مجموعه دادگان EYEDIAP



شکل (۲-۵) مقایسه دقت روش‌های مختلف بر روی مجموعه دادگان MPIIGaze

در اینجا به منظور سهولت ابتدا آزمایش‌های مختلف را به همراه کارهای مرتبط اخیر به طور خلاصه بیان می‌کنیم.

آزمایش شماره ۱: تشخیص جهت نگاه با استفاده از اطلاعات تصویر صورت و جهت سر

آزمایش شماره ۲: تشخیص جهت نگاه با استفاده از اطلاعات استخراج شده به روش عمیق از روی جهت سر

آزمایش شماره ۳: تشخیص جهت نگاه با استفاده از اطلاعات جهت سر- بهبود دقت

آزمایش شماره ۴: تاثیر افزایش تعداد لایه‌ها بر دقت خروجی

آزمایش شماره ۵: استفاده از رگرسیون خطی به عنوان دسته‌بند

آزمایش شماره ۶: استفاده از ماشین بردار پشتیبان رگرسیون به عنوان دسته‌بند

[۲۹] تشخیص جهت نگاه از روی ظاهر در محیط طبیعی

[۳۰] تشخیص جهت نگاه برای همه

[۳۱] تشخیص جهت نگاه از روی صورت- استفاده از وزن‌های مکانی با پیچش‌های 1×1

در شکل‌های (۲-۵) و (۱-۵) از سمت چپ به ترتیب آزمایش‌های ۱ تا ۶، سامانه پیشنهادی و دقت‌های بهدست‌آمده از [۲۹]، [۳۰] و [۳۱] نشان داده شده‌اند.

همان‌طور که در شکل (۱-۵) و شکل (۲-۵) مشاهده می‌شود بهترین دقت مربوط به آزمایش شماره ۱ است. هدف این آزمایش تشخیص جهت نگاه با استفاده از تصاویر برش خورده چشم‌ها و به کمک اطلاعات

جهت سر است. در این آزمایش اطلاعات جهت سر به‌طور مستقیم از مجموعه دادگان وارد مدل شده‌اند. پس از این آزمایش، بهترین دقت مربوط به شبکه ارائه شده در [۳۱] است. در این شبکه از مدل وزن‌های مکانی برای اجزا صورت استفاده شده است. آن‌ها تنها از اطلاعات موجود در تصویر صورت استفاده کرده و خطای ۶ و ۴,۸ درجه‌ای را به ترتیب برای مجموعه‌های دادگان EYEDIAP و MPIIGaze گزارش کرده‌اند. ایده ارائه شده در این مقاله که در زمان نوشتن این پایان‌نامه تنها ۲ ماه قدمت دارد، بسیار شبیه مدل پیشنهادی این پایان‌نامه است. مدل پیشنهادی با اندکی اختلاف و با خطای ۷,۵ و ۶,۴ درجه‌ای در جایگاه سوم قرار دارد. مزیت این مدل همچون [۳۱] نسبت به آزمایش شماره ۱ این است که به جز فریم‌های ورودی به هیچ‌گونه اطلاعاتی از مجموعه دادگان نیازی نداشته و کاملاً انتهایاً به انتها عمل می‌کند. ما نیز در روش پیشنهادی بر این اعتقاد هستیم علیرغم این که به لحاظ نظری، تفاوت مشارکت بخش‌های مختلف صورت در تخمین جهت نگاه می‌تواند توسط شبکه عصبی پیچشی آموخته شود، ارائه مدلی که شبکه را در آموزش بهتر این تفاوت مجبور کند حائز اهمیت است. مدل گرافیکی ارائه شده همانند [۳۱] وابستگی مکانی اجزای صورت را در نظر گرفته و تفاوت مشارکت بخش‌های مختلف صورت را در تخمین جهت نگاه انسان لحاظ می‌کند.

آزمایش شماره ۲ که به ترتیب ۱۲ و ۱۰,۳ درجه خطا را به دست داده است مربوط به تخمین ماتریس‌های چرخش و انتقال برای جهت سر با استفاده از یادگیری عمیق (LeNet) است. این ماتریس‌ها پس از محاسبه در مدلی که در آزمایش ۱ استفاده شد، به عنوان اطلاعات سر مجددًا جایگذاری شده و مدل دوباره فرآیند آموزش را طی می‌کند. خطای به دست آمده نسبتاً بالا بوده و دلیل این امر استفاده کاملاً انتهایاً در یادگیری جهت سر است. همان‌طور که قبلًاً هم در فصل ۴ توضیح داده شد، برای تخمین جهت سر تنها از برچسب ماتریس‌های چرخش و انتقال، بدون هیچ‌گونه پیش‌پردازشی استفاده شده است. در این آزمایش تنها از نیمی از دادگان استفاده شده است.

در آزمایش ۳ به منظور بهبود دقت، شبکه استفاده شده در آزمایش ۲ را تغییر داده‌ایم. این روند شامل تغییر در اندازه لایه‌های مختلف پیچش بوده است. سرانجام افزودن یک لایه پیچش با خروجی ۱ بهترین نتیجه را در این آزمایش شامل دقت‌های ۹ و ۸ درجه به دست داده است.

هدف از آزمایش ۴، تغییر شبکه عمیق استفاده شده در آزمایش‌های قبل در جهت افزایش تعداد لایه‌های پیچش بوده است. در همین راستا از شبکه VGG-16 استفاده کرده‌ایم. از آنجا که در این آزمایش‌های از فریم‌های بریده شده تصویر چشم به عنوان ورودی اولیه استفاده شده است و این فریم‌ها پیچیدگی‌های هندسی زیادی ندارند، نتایج این آزمایش (خطای ۱۱ و ۱۰ درجه‌ای) نشان می‌دهد افزایش تعداد لایه‌های پیچش که به معنای افزایش تعداد فیلترهای مختلف است کمکی به بهبود نتیجه نمی‌کند.

در آزمایش ۵ و ۶ هدف، ارزیابی ویژگی‌های استخراج شده توسط زیر بخش یادگیری عمیق در معماری پیشنهادی است. همان‌طور که شکل‌های (۱-۵) و (۲-۵) مشاهده می‌شود، ویژگی‌های استخراج شده در این مرحله که تنها از صورت استخراج شده و هیچ‌گونه بر Shi در چشم‌ها و یا استفاده از جهت سر دیده نمی‌شود، موجب بهبود نتایج شده است. به علاوه استفاده از ماشین بردار رگرسیون که دسته‌بندی قوی‌تری نسبت به رگرسیون خطی است نشان می‌دهد بهره‌گیری از یک دسته‌بند قوی‌تر می‌تواند تفکیک‌پذیری بین ویژگی‌های استخراج شده را بهتر انجام دهد.

در نهایت در سامانه پیشنهادی و در ادامه روند دو آزمایش قبل، از دسته‌بند قوی‌تری استفاده شده است. در این آزمایش که از ترکیبی از استخراج ویژگی توسط شبکه AlexNet و مدل گرافیکی میدان‌های شرطی MPIIGaze استفاده شده، به ترتیب خطای ۷,۵ و ۶,۴ درجه در مجموعه‌های دادگان EYEDIAP و MPIIGaze به دست آمده‌اند. این میزان خطای نشان می‌دهد نتیجه‌گیری به دست آمده از آزمایش‌های ۵ و ۶ صحیح بوده و با افزایش قدرت دسته‌بند، که توانایی زیادی در مدل‌سازی ارتباط‌های مکانی درون فریمی (و البته مدل‌سازی زمانی بین فریمی) دارد نتایج بهتری حاصل شده‌اند.

خطاهای گزارش شده در مقالات و خطاهای به دست آمده در آزمایش‌های مختلف به دلیل این که جهت سر در مجموعه دادگان MPIIGaze نسبتاً دارای پیچیدگی کمتری به نسبت EYEDIAP است مقادیر کمتری را به خود اختصاص داده‌اند.

در مورد آموزش مدل گرافیکی احتمالاتی ارائه شده در معماری پیشنهادشده این نکته حائز اهمیت است که این مدل بر روی CPU قابلیت آموزش دارد. از این‌رو با توجه به این که در این مدل پارامترهای قابل تنظیم مختلفی وجود دارند و نیز نظر به محدودیت‌های سخت‌افزاری فرآیند آموزش بسیار زمان‌بر بوده است. به نظر می‌رسد با افزایش توان محاسباتی و تغییر پارامترهای مختلف این مدل بتواند تفاوت مشارکت بخش‌های مختلف صورت را در تخمین جهت نگاه انسان به خوبی مدل کرده و بهترین نتایج را از آن خود کند.

فصل ۶:

جمع‌بندی و پیشنهادها

۱-۶- جمع‌بندی

در این پایان‌نامه معماری جدیدی برای تشخیص جهت نگاه انسان ارائه شده است. این معماری بر پایه دو بخش مهم یادگیری ماشین یعنی یادگیری عمیق و مدل‌های گرافیکی احتمالاتی بیان شده است. در ابتدا از چارچوب یادگیری عمیق به منظور استخراج ویژگی‌های مختلف استفاده شده است. در همین راستا آزمایش‌های متعددی انجام شده است تا دقیقت خروجی به ازای شبکه‌های مختلف یادگیری عمیق با ویژگی‌های مختلف بررسی شوند. درنهایت در قالب معماری پیشنهادی از یک شبکه یادگیری عمیق مبتنی بر AlexNet به منظور استخراج ویژگی از سه عضو مهم صورت یعنی چشم‌ها، دهان و بینی استفاده شده است. ویژگی‌های استخراج شده به یک مدل گرافیکی احتمالاتی با نام میدان‌های تصادفی شرطی پنهان داده شده‌اند تا خروجی نهایی در قالب یک مختصات سه‌بعدی از جهت نگاه انسان حاصل شود. نتایج به دست آمده نشان می‌دهند ویژگی‌های استخراج شده ویژگی‌هایی مفید در تشخیص جهت نگاه انسان بوده و مدل میدان‌های شرطی تصادفی پنهان نیز توانایی بالایی در استفاده از این ویژگی‌ها در تخمین نهایی نقطه‌ای که انسان به آن می‌نگرد دارد.

با توجه به چالش‌هایی که در مسیر تحقیقات و پیاده‌سازی این سامانه به وجود آمده و راه حل‌هایی که برای آن‌ها در نظر گرفته شد، می‌توان مواردی را به عنوان موارد تحقیقاتی آتی در حوزه سامانه تشخیص جهت نگاه انسان ارائه کرد که در بخش بعدی بیان شده‌اند.

۲-۶- پیشنهادها

محدودیت‌های مختلفی در مسیر تحقیقاتی سامانه پیشنهادی وجود داشته است. در حوزه سخت‌افزاری به دلیل محدودیت‌هایی که در منابع سیستمی از جمله قدرت پردازنده گرافیکی وجود داشت، امکان بررسی شبکه‌های جدیدی همچون DeepID3 [۶۴] که در بازناسی صورت استفاده می‌شوند وجود نداشته است. پیشنهاد می‌شود تأثیر استفاده از شبکه‌هایی همچون DeepID3 بر روی استخراج ویژگی‌های عمیق بررسی شوند.

در این پایان‌نامه به مدل‌سازی و تشخیص جهت چشم در یک فریم پرداخته شده است به طوری که گره‌های ویژگی همه مربوط به تصویر صورت در یک فریم بوده‌اند. پیشنهاد دیگر مدل‌سازی تشخیص جهت

نگاه انسان در ویدئو به کمک مدل گرافیکی احتمالاتی است به طوری که یک گره نمایانگر ویژگی‌های استخراج شده در یک فریم و گره‌های دیگر نماینده ویژگی‌های فریم‌های دیگر باشند. بدین ترتیب می‌توان با ساختار گراف زنجیره تأثیر ارتباط زمانی را در ویدئو بر تخمین جهت نگاه انسان بررسی کرد.

یکی از مباحثی که اخیراً در حوزه یادگیری عمیق توجه زیادی را به خود جلب کرده است شبکه‌های رقابتی تولیدکننده^۱ است. [۶۵] این شبکه‌های GAN که بر آموزش بدون ناظر تأکید دارند توانسته‌اند نتایج شگرفی را در بخش‌های مختلفی از جمله تصاویر صورت [۶۶] از خود نشان دهند. اخیراً از این شبکه‌ها در تولید تصاویر چشمی نیز استفاده شده است [۶۷]. به نظر می‌رسد در مسئله تشخیص جهت نگاه نیز بتوان از یادگیری بدون ناظر در قالب شبکه‌های رقابتی تولیدی استفاده کرد. پیشنهاد می‌شود تلفیق این شبکه‌های رقابتی تولیدی با تشخیص جهت نگاه مورد بررسی قرار گیرد.

¹ Generative Adversarial Networks

مراجع

مراجع

- [۱] S. Ghosh, T. Nandy, and N. Manna, "Real Time Eye Detection and Tracking Method for Driver Assistance System," in *Advancements of Medical Electronics*, ed: Springer, 2015, pp. 13-25.
- [۲] M. R. Wilson, J. S. McGrath, S. J. Vine, J. Brewer, D. Defriend, and R. S. Masters, "Perceptual impairment and psychomotor control in virtual laparoscopic surgery," *Surgical endoscopy*, vol. 25, pp. 2268-2274, 2011.
- [۳] S. Instruments. (2016). *SPORTS, PROFESSIONAL TRAINING*. Available: <http://www.smivision.com/en/gaze-and-eye-tracking-systems/applications/sports-professional-training-education.html>
- [۴] Hansen, D. Witzner, and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* pp. 478-5. ۲۰۱۰ , .
- [۵] Zhu, Zhiwei, K. Fujimura, and Q. Ji, "Real-time eye detection and tracking under various light conditions," *Proceedings of the 2002 symposium on Eye tracking research & applications. ACM,*, 2002.
- [۶] Valenti, Roberto, and T. Gevers, "Accurate eye center location and tracking using isophote curvature," *Computer Vision and Pattern Recognition*, 2008.
- [۷] Hansen, D. Witzner, and A. E. Pece, "Eye tracking in the wild," *Computer Vision and Image Understanding*, vol. 98, pp. 155-181, 2005.
- [۸] C. Choo, J. W. Lee, K. Y. Shin, E. C. Lee, K. R. Park, H. Lee, *et al.*, "Gaze Detection by Wearable Eye-Tracking and NIR LED-Based Head-Tracking Device Based on SVR. ET," *Etri Journal*, 2012.
- [۹] Yuille, A. L., P. W. Hallinan, and D. S. Cohe, "Feature extraction from faces using deformable templates," *International journal of computer vision*, vol. 8, pp. 99-111, 1992.
- [۱۰] I. F. Ince and J. W. Kim, "A 2D Eye Gaze Estimation System With Low Resolution Webcam Images," *EURASIP Journal on Advances in Signal Processing*, 2012.
- [۱۱] J. Waite and J. M. Vincent, "A probabilistic framework for neural network facial feature location," *British Telecom Technology Journal*, vol. 10, 1992.
- [۱۲] J. Bala, K. DeJong, J. Huang, H. Vafaie, and H. Wechsler, "Visual routine for eye detection using hybrid genetic architectures," *Pattern Recognition*, 1996., *Proceedings of the 13th International Conference on*, vol. 3, 1996.
- [۱۳] Kawato, Shinjiro, and J. Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the between-eyes," *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000.
- [۱۴] L. Orazio T. D., M., G. Cicirelli, and A. Distante, "An algorithm for real time eye detection in face images," *ICPR 2004 Proceedings of the 17th International Conference on*, IEEE, vol. 3, 2004.
- [۱۵] M. Wilson, J. McGrath, S. Vine, J. Brewer, D. Defriend, and R. Masters, "Perceptual impairment and psychomotor control in virtual laparoscopic surgery," *Springer*, 2011.

- [۱۶] P. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi, "Real-Time Eye Gaze Tracking for Gaming Design and Consumer Electronics Systems," *IEEE Trans. On Consumer Electronics*, vol. 58, 2012.
- [۱۷] Lewis and J. R, "In the eye of the beholder," *IEEE Spectrum*, vol , ۲۰ .pp. 24-28, 2004.
- [۱۸] Hallinan and P. W, "Recognizing human eyes," *International Society for Optics and Photonics*, 1991.
- [۱۹] Viola, Paul, and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, 2004.
- [۲۰] Ishikawa and Takahiro, "Passive driver gaze tracking with active appearance models," *In Proceedings of the 11th World Congress on Intelligent Transportation Systems*, 2004.
- [۲۱] D. W. Hansen, J. P. Hansen, M. Nielsen, A. S. Johansen, and M. B. Stegmann" ,Eye typing using markov and active appearance models," *In IEEE Workshop on Applications on Computer Vision*, pp. 132-136, 2003.
- [۲۲] T. F. Cootes, G. J. Edwards, and C. J. Taylor., "Active appearance models," *In Proc. European Conf. on Computer Vision*, Springer, vol. 2, 1998.
- [۲۳] D. W. Hansen, "Using Colors for Eye Tracking,,," *chapter Color Image Processing: Emerging Applications*, pp. 309-327.
- [۲۴] N. Markuš, M. Frljak, I. S. Pandžić, J. Ahlberg, and R. Forchheimer, "Eye pupil localization with an ensemble of randomized trees," *Pattern recognition* vol. 47, 2014.
- [۲۵] K. Alberto, F. Mora, J.-m. Odobez, and D. Lausanne, "Geometric Generative Gaze Estimation (G 3 E) for Remote RGB-D Cameras," *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.
- [۲۶] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," *Proceedings of the IEEE International Conference on Computer Vision*, 2013.
- [۲۷] Y.-m. Cheung, S. Member, and Q. Peng, "Eye Gaze Tracking With a Web Camera in a Desktop Environment," 2015.
- [۲۸] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," *Proceedings of the IEEE International Conference on Computer Vision*, 2011.
- [۲۹] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4511-4520.
- [۳۰] K. Kafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, *et al.*, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2176-2184.
- [۳۱] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation," *arXiv preprint arXiv:1611.08860*, 2016.
- [۳۲] F. Rosenbaltt, "The Perceptron—a Perciving and Recognizing Automation," Report 85-460-1 Cornell Aeronautical Laboratory, Ithaca1957.

- [۳۳] W. Hu, D. Xie, T. Tan, and S. Maybank, "Learning activity patterns using fuzzy self-organizing neural network," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, pp. 1618-1626, 2004.
- [۳۴] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, pp. 1527-1554, 2006.
- [۳۵] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, pp. 1-127, 2009.
- [۳۶] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.
- [۳۷] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 609-616.
- [۳۸] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Hoseini, and M. Fathy, "Online signature verification based on feature representation," in *Artificial Intelligence and Signal Processing (AISP), 2015 International Symposium on*, 2015, pp. 211-216.
- [۳۹] L. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, 2001.
- [۴۰] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 1036-1043.
- [۴۱] .(۱۳۹۰/۰۶/۱۶)Feature extraction using convolution - Ufldl. Available: http://ufldl.stanford.edu/wiki/index.php/Feature_extraction_using_convolution
- [۴۲] .(۱۳۹۰/۰۶/۱۶)Pooling - Ufldl. Available: <http://ufldl.stanford.edu/wiki/index.php/Pooling>
- [۴۳] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
- [۴۴] S. Nowozin and C. H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, pp. 185-365, 2011.
- [۴۵] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, pp. 93-128, 2006.
- [۴۶] D. Koller, N. Friedman, L. Getoor, and B. Taskar, "Graphical models in a nutshell," *Introduction to statistical relational learning*, pp. 13-55, 2007.
- [۴۷] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, et al., "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675-678.
- [۴۸] S. H. Hassanzadeh. (2016). *Caffe Training*. Available: <http://deeplearning.ir/>
- [۴۹] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
- [۵۰] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, 2011.

- [۰۱] N. Andy. *AN EXPLANATION OF XAVIER INITIALIZATION*. Available: <http://andyljones.tumblr.com/post/110998971763/an-explanation-of-xavier-initialization>
- [۰۲] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [۰۳] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, 2012.
- [۰۴] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, *et al.*, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural networks: the statistical mechanics perspective*, vol. 261, p. 276, 1995.
- [۰۵] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278-2324, 1998.
- [۰۶] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Aistats*, 2010, pp. 249-256.
- [۰۷] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [۰۸] D. Frossard. (2016). *VGG in TensorFlow*. Available: <https://www.cs.toronto.edu/~frossard/post/vgg16/>
- [۰۹] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3676-3684.
- [۱۰] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
- [۱۱] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, 2007.
- [۱۲] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Distributed structured prediction for big data," in *NIPS workshop on Big Learning*, 2012.
- [۱۳] C.-T. Yuan and S. Chen, "Message Passing Interface (MPI)," in *Euro-Par'96 Parallel Processing*, 1996, pp. 128-135.
- [۱۴] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.
- [۱۵] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu ,D. Warde-Farley, S. Ozair, *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.
- [۱۶] H. Cate, F. Dalvi, and Z. Hussain, "DeepFace: Face Generation using Deep Learning," *arXiv preprint arXiv:1702.01876*.
- [۱۷] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 131-138.
- [۱۸] K. A. F. Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d

cameras," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2014, pp. 255-258.

پیوست‌ها

مجموعه‌های دادگان^{۱۳۷}

در این پژوهش از دو مجموعه دادگان استفاده شده است:

- مجموعه دادگان EYEDIAP[68]

- مجموعه دادگان MPIIGaze[29]

مجموعه دادگان EYEDIAP

این مجموعه دادگان که در اواخر سال ۲۰۱۴ منتشر شده است شامل فایل‌های ویدئویی با طول حدود ۴ دقیقه از ۱۶ شرکت‌کننده است. این فایل‌های ویدئویی به صورت ویدئوهای عمق^{۱۳۸}، ویدئوهای باوضوح VGA و همچنین ویدئوهای با کیفیت HD ضبط شده‌اند. شرکت‌کنندگان در دو حالت کلی به دو هدف نگاه می‌کنند. اهداف موردنظر یا به صورت سه‌بعدی در فضا حرکت می‌کنند (یک توب) و یا بر روی صفحه‌نمایش قرار دارند. اهداف روی صفحه‌نمایش خود به صورت گسسته^{۱۳۹} و یا پیوسته در حال حرکت کنند. در هر کدام از این حالات، فرد در دو حالت کلی، یا با سر کاملاً ثابت هدف موردنظر را می‌نگرد و یا به‌طور آزاد سر خود را حرکت می‌دهد. جدول (۲-۵) خلاصه‌ای از این حالات مختلف ضبط ویدئو را نشان می‌دهد.

جدول (۱-۶) خلاصه وضعیت جلسات ضبط ویدئو در مجموعه دادگان [۶۸]

Participants	Recorded sessions (the participant index is implicit)
1-11	A-DS-S; A-DS-M; A-CS-S; A-CS-M; A-FT-S; A-FT-M
12-13	B-FT-S; B-FT-M
14-16	A-DS-S; A-DS-M; A-CS-S; A-CS-M; A-FT-S; A-FT-M B-FT-S; B-FT-M

برای مثال، ۱_A_DS_S مشخص‌کننده شرکت‌کننده شماره ۱ است که با حالت سر ثابت (S) به اهداف

۱۳۷ Datasets

۱۳۸ depth

۱۳۹ discrete

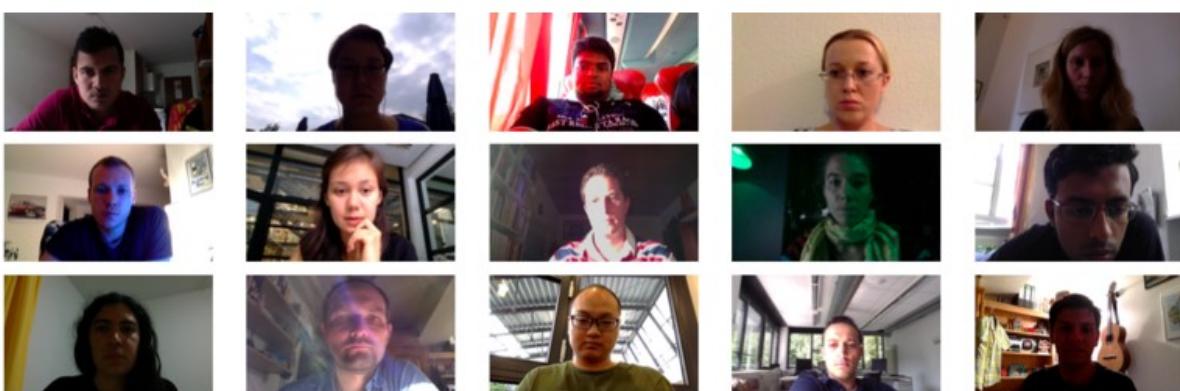
گسسته بر روی نمایشگر (DS) نگاه کرده است. ۵ نیز نمایانگر شرکت‌کننده شماره ۵ است که با حالت سر دارای حرکت (M) به توپی که در فضا در حال حرکت بوده (FT) نگاه کرده است. شکل (۱-۶) نمونه‌هایی از فریم‌های این مجموعه دادگان را نشان می‌دهد.



شکل (۱-۶) نمونه‌هایی از فریم‌های ضبط شده در مجموعه دادگان EYEDIAP [۶۸]

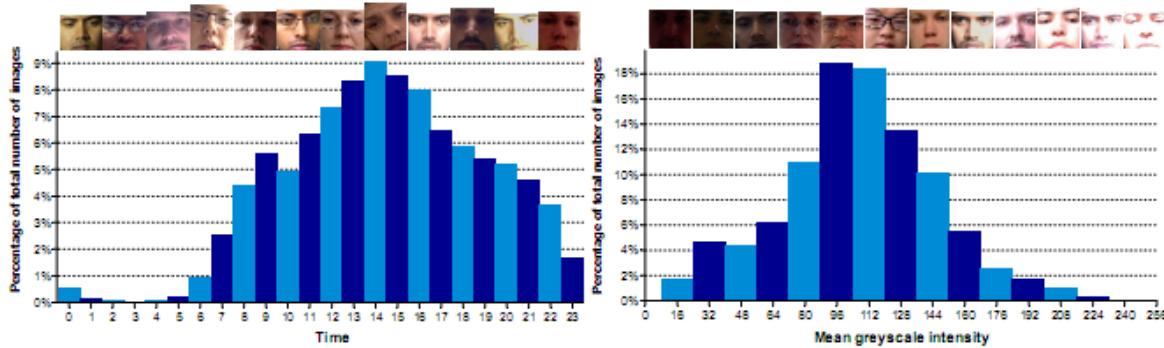
در این مجموعه دادگان علاوه بر فایل‌های ویدئویی ذکر شده اطلاعات زیر نیز موجود هستند:

- حالت سر بیننده در هر فریم بر اساس ماتریس‌های چرخش و انتقال
 - موقعیت مکانی دو و سه‌بعدی عددی نقطه‌ای که کاربر بر روی نمایشگر و یا در فضای آن نگاه می‌کند.
 - موقعیت مکانی مرکز مردمک چشم در هر فریم بر حسب پیکسل
- ### MPIIGaze مجموعه دادگان
- در این مجموعه دادگان که در سال ۲۰۱۵ منتشر شده است مجموعاً ۲۱۳۶۵۹ تصویر از ۱۵ شرکت‌کننده وجود دارد. تعداد تصاویری که از هر شرکت‌کننده گرفته شده است بین ۳۴۷۴۵ تا ۱۴۹۸ مورد متفاوت است. نمونه‌هایی از تصاویر این مجموعه دادگان را در شکل (۲-۶) مشاهده می‌کنید.



[۲۹] نمونه‌هایی از تصاویر موجود در مجموعه دادگان MPIIGaze

شکل (۲-۶) چپ نیز، پراکندگی جلسات دریافت تصاویر را در طول ساعات مختلف شباهه‌روز (۰ تا ۲۴) نشان می‌دهد. شکل (۳-۶) سمت راست نیز پراکندگی شدت روشنایی را در این مجموعه دادگان نشان می‌دهد.



[۲۹] برخی مشخصات مجموعه دادگان MPIIGaze

در فرآداده^{۱۴۰}‌های همراه با این مجموعه دادگان اطلاعات زیر قابل استخراج هستند:

- موقعیت نشانه‌های چشمی^{۱۴۱} بر حسب پیکسل در تصویر اصلی دریافت شده از شرکت‌کننده
- موقعیت مکانی دو بعدی نقطه‌ای که کاربر بر روی نمایشگر به آن نگاه می‌کند.
- موقعیت سه بعدی نسبت به دوربین برای نقطه‌ای که کاربر بر روی نمایشگر به آن نگاه می‌کند
- موقعیت سر شرکت‌کننده بر حسب ماتریس‌های چرخش و انتقال

140 metadata

141 Eye landmarks

Abstract:

Human Gaze Estimation consists of Eye tracking and providing the computational model for gaze estimation. Human gaze estimation plays a crucial role in expressing a person's desires, needs, cognitive processes, emotional states, and interpersonal relations. The importance of eye movements to the individual's perception of and attention to the visual world is implicitly acknowledged, as it is the method through which we gather the information necessary to negotiate our way through and identify the properties of the visual world. Human gaze estimation has many applications in behavior and attention analysis, human-computer interaction, etc.

There have been numerous methods for tracking eyes and estimating gaze, but despite active research and significant progress in the last 20 years, gaze estimation remains challenging due to the individuality of eyes, occlusion, variability in scale and head pose rotation, location, and light conditions. In this research, we have investigated recent methods and presented a new architecture, based on convolutional neural networks and probabilistic graphical models. We have used EYEDIAP and MPIIGaze datasets and did multiple experiments. The results show Mean Error degree of 7.5 and 6.4 for mentioned datasets, respectively.

Keywords:

Human Gaze Estimation, Eye Tracking, Attention, Behavior, Convolutional Neural Networks, Probabilistic Graphical Models



**Iran University of Science and Technology
Computer Engineering Department**

Human Gaze Estimation Using Probabilistic Graphical Models

**A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree
of Master of Science (Doctor of Philosophy) in Computer Engineering**

**By:
Rahim Entezari**

**Supervisor:
Dr. Mahmood Fathy
Dr. Reza Berangi**

February 2017