نشریه علمی مهندسی فناوری اطلاعات مکانی

سال سیزدهم، شماره اول، بهار ۱۴۰۴

Vol.13, No.1, Spring 2025

۱ -۱۷

مقاله پژوهشی

پیش‌بینی موقعیت سه‌بعدی عابر پیاده با استفاده از یادگیری عمیق از روی داده‌های کینکت

اکبر جعفری[۱]، علی حسینی نوه[۲]*، مجتبی محمودیان[۳]

۱- دانشجوی دکتری، گروه فتوگرامتری و سنجش از دور، دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه نصیرالدین طوسی

۲- دانشیار، گروه فتوگرامتری و سنجش از دور، دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه نصیرالدین طوسی

۳- دانشیار، گروه مهندسی عمران و زیرساخت‌ها، دانشکده مهندسی، دانشگاه *RMIT* ملبورن، استرالیا

تاریخ دریافت مقاله: ۱۴۰۲/۰۳/۲۰ تاریخ پذیرش مقاله: ۱۴۰۳/۰۳/۰۶

چکیده

پیش‌بینی مسیر حرکت عابر پیاده از موضوعات مهم در حوزه بینایی ماشین و سامانه‌های حمل‌ونقل هوشمند است، زیرا بر ایمنی و قابلیت تصمیم‌گیری سیستم‌های خودران تأثیر مستقیم دارد. اغلب رویکردهای موجود با استفاده از داده‌های دوبعدی (*RGB*) و شبکه‌های بازگشتی نظیر *LSTM (Long Short Term Memory)* توسعه یافته‌اند، اما این روش‌ها بعد عمق را نادیده گرفته و در نتیجه برآورد فاصله میان عابران و عوارض پیرامونی به‌دقت انجام نمی‌شود. در این پژوهش، یک مدل (*Three Dimension- LSTM*) *3D-LSTM* معرفی می‌شود که با استفاده از داده‌های *RGB-D* حاصل از حسگر *Kinect* ثابت، پیش‌بینی موقعیت عابران پیاده را در فضای سه‌بعدی متریک انجام می‌دهد. فرآیند مدل‌سازی شامل استخراج داده‌های عمق از تصاویر استریو، نرمال‌سازی مختصات و آموزش شبکه *LSTM* برای پیش‌بینی مختصات سه‌بعدی (*X, Y, Z*) در گام‌های آینده است. نتایج حاصل از ارزیابی بر روی مجموعه داده دانشگاه پلی‌تکنیک لوزان (*EPFL*) نشان می‌دهد که میانگین دقت پیش‌بینی سه‌بعدی ۱۵٫۷۰سانتیمتر(تقریباً معادل روش‌های دوبعدی) است، اما در عین حال اطلاعات فاصله واقعی و تعاملات فضایی را نیز در خروجی ارائه می‌دهد که برای جلوگیری از برخورد و برنامه‌ریزی مسیر بسیار حیاتی است. تحلیل‌ها حاکی از آن است که افزودن بعد سوم نه‌تنها باعث افت عملکرد نمی‌شود، بلکه منجر به بهبود قابلیت تصمیم‌گیری در شرایط واقعی می‌گردد. این روش می‌تواند مبنایی برای توسعه سیستم‌های هوشمند ناوبری، رباتیک و خودروهای خودران با درک دقیق‌تر از محیط سه‌بعدی باشد.

کلیدواژه‌ها : پیش‌بینی موقعیت سه‌بعدی، شبکه‌های حافظه کوتاه و بلندمدت ، عابر پیاده، یادگیری عمیق.

* نویسنده مکاتبه‌کننده: تهران، خیابان ولیعصر، بالاتر از میدان ونک، تقاطع میرداماد، دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه نصیرالدین طوسی.

تلفن: ۰۲۱-۸۸۸۸۲۹۹۱

Email: hosseininaveh@kntu.ac.ir

[DOI: 10.61882/jgit.13.1.1]

۱- مقدمه

امروزه پیش‌بینی مسیر حرکت عابر پیاده، یکی از چالش‌های مهم در حوزه ماشین‌بینایی بوده و توجه پژوهشگران زیادی را به خود جلب نموده است. از کاربردهای مهم پیش‌بینی مسیر حرکت عابر پیاده می‌توان به سیستم‌های هدایت و ناوبری رباتها و خودروهای هوشمند اشاره نمود[۱و۳]. در اغلب موارد پیش‌بینی مسیر حرکت انسان امری مشکل می‌باشد، زیرا در هر لحظه ممکن است فرد، تغییر جهت داده و یا توقف نماید[۴]. فرایند پیش‌بینی مسیر حرکت عوارض دینامیک به ویژه عابر پیاده به عوامل مختلف، مانند سرعت حرکت، هدف عابر پیاده، موانع موجود در مسیر حرکت، پیاده‌روی فردی یا گروهی و میزان رعایت قوانین ترافیکی توسط اشخاص وابسته است[۵و۶]. در حالت کلی برای پیش‌بینی مسیر انسان باید مسیر حرکت گذشته او مطالعه شده و بر مبنای آن موقعیت آینده‌اش پیش‌بینی شود[۷]. مطالعات نشان می‌دهد که الگوی حرکتی انسانها، وجود موانع در مسیر و واکنشهای انسانی بیشترین تأثیر را در مسیر حرکتی انسان دارد[۸و ۹].

در گذشته، برای پیش‌بینی مسیر حرکت انسانها، از الگوریتم‌هایی مانند کالمن فیلتر[۱۰] فیلتر اجزاء[1] [۱۱] و پردازش گاوسی[2][۱۲] استفاده می‌شد. این روش‌ها، موقعیت در زمان آتی را بر مبنای وضعیت فعلی پیش‌بینی می‌کنند و نمی‌توانند اطلاعات زیادی از مسیرها و الگوی حرکتی انسان را در حافظه خود برای پیش‌بینی موقعیت آتی نگهداری کنند[۷].

امروزه برای مسائلی که به پیش‌بینی نیاز دارند از الگوریتم‌های مطرح یادگیری عمیق استفاده می‌شود[۱۳]. برای پیش‌بینی مواردی مانند مسیر حرکت عوارض متحرک مثل عابرین پیاده یا خودرو باید از شبکه‌های بازگشتی که قابلیت حفظ اطلاعات لازم از مسیر گذشته را دارند، استفاده نمود[۱۴]. یکی از معایب شبکه‌های بازگشتی این است که قادر به پشتیبانی از دنباله‌های بسیار طولانی نیستند و محدود به چند گام قبل هستند[۱۵]. در حالی که در بسیاری از مسائل پیش‌بینی، نیاز به حفظ اطلاعات برای مدت طولانی می‌باشد. به عبارت دیگر برای پیش‌بینی آینده، اطلاعات زمانهای چندین گام قبل مورد نیاز می‌باشد. شبکه‌های حافظه کوتاه و بلندمدت[3]*(LSTM)* که نوعی از شبکه‌های بازگشتی[4] *(RNN)* هستند، توانایی حفظ اطلاعات برای مدت طولانی را دارند[۱۶]. این شبکه‌ها علیرغم اینکه در ابتدا برای پیش‌بینی کلمات بعدی در جملات استفاده شدند[۱۷]، ولی محققین در پیش‌بینی مسیر عابرین پیاده با استفاده از *LSTM* نیز نتایج خیلی خوبی بدست آوردند[۱۸و ۱۹].

شبکه‌های *LSTM* مختلفی برای پیش‌بینی مسیر عابرین پیشنهاد شده است. از جمله می توان به شبکه *LSTM* اجتماعی[5]*(S-LSTM)* اشاره نمود که علاوه بر مدل اصلی *LSTM*، تأثیر عوارض نزدیک به عابر را نیز در مدل‌سازی و پیش‌بینی مسیر حرکت در نظر می‌گیرد[۱۹و۲۰].

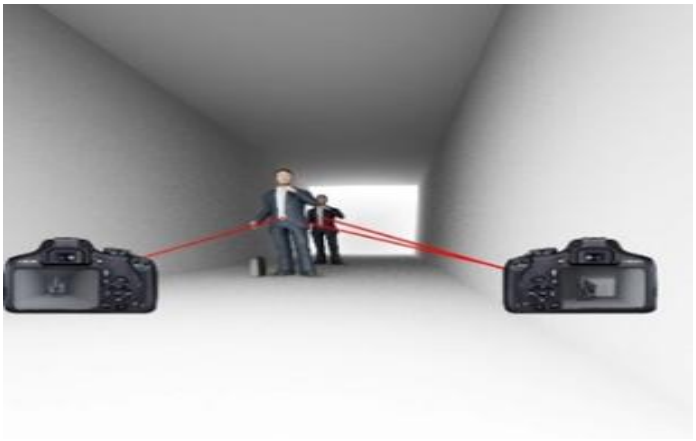
در شبکه‌های پیشنهادی *LSTM* برای پیش‌بینی مسیر عابرین با استفاده از داده‌های تصویری، معمولاً داده‌های تصویری شامل سه باند سبز-قرمز- آبی[6]*(RGB)* بعد از نرمالسازی استفاده می‌شود[۱۸]. از آنجاییکه داده‌های *RGB* در فضای دوبعدی و پیکسل مبنا می‌باشند، لذا شبکه‌های پیشنهادی *LSTM* پیش‌بینی مسیر حرکت را در فضای دو بعدی انجام می‌دهند[۱۸و۲۱]. برای پیش‌بینی بهتر وضعیت عابر پیاده لازم است مدل پیشنهادی به دنیای واقعی نزدیک‌تر باشد. بدین منظور باید عوامل تأثیرگذار در رفتار انسان را شناسایی نموده

[1] *Particle Filter*
[2] *Gaussian Processes*
[3] *Long-Short Term Memory*
[4] *Recurrent Neural Network*
[5] *Social LSTM*
[6] *Red-Green-Blue*

[DOI: 10.61882/jgit.13.1.1]

و در تصمیم‌سازی و پیش‌بینی وضعیت عابر پیاده دخالت داده شوند. با توجه به اینکه دنیای واقعی سه‌بعدی است و یکی از مهمترین عوامل در رفتار عابر پیاده فاصله مابین انسان و عوارض دیگر، بویژه عوارض متحرک می‌باشد، بنابراین باید پیش‌بینی موقعیت آتی نیز در فضای سه‌بعدی انجام گیرد. ولی تابحال محقّقین در مدل‌های پیشنهادی برای پیش‌بینی موقعیت آتی، بر مبنای فضای دوبعدی مدلسازی نموده‌اند.

به عبارت دیگر با تصویر فضای سه‌بعدی دنیای واقعی به فضای دوبعدی در حالت تک‌تصویر، داده‌های تهیه شده دارای نقصان اطلاعات بوده و از بعد سوم یا همان عمق، اطلاعاتی را ارائه نمی‌دهد. همان طور که در شکل (۱) ملاحظه می‌شود، اگر همزمان با زوج دوربین از فضای سه‌بعدی عکس تهیه شود و با استفاده از تقاطع پرتو نور متناظر موقعیت سه‌بعدی عابر پیاده استخراج شود، می‌توان علاوه بر پیش‌بینی موقعیت دوبعدی، بعد سوم یا همان عمق را نیز پیش بینی نمود.



شکل۱: تهیه زوج تصاویر متوالی از عابر پیاده در حال حرکت

لذا در این تحقیق از داده‌های تصویری همراه با عمق[1] *(RGB-D)* برای مدلسازی شبکه *LSTM* استفاده شده است تا تأثیر فاصله عابر پیاده با دیگر عوارض در فضای سه‌بعدی در پیش‌بینی موقعیت عابرین نشان داده شود. هدف از این تحقیق، ارزیابی پیش‌بینی موقعیت سه‌بعدی عابر پیاده بر مبنای داده‌های *RGB-D* با استفاده از شبکه *LSTM* و مقایسه نتایج آن با حالت دوبعدی می‌باشد.

در بخش دوم مروری بر کارهای مرتبط با موضوع تحقیق انجام گرفته است. در بخش سوم روش پیشنهادی و مراحل انجام تحقیق توضیح داده شده است. در بخش چهارم نتایج بدست‌آمده و دقت پیش‌بینی توضیح داده شده و در بخش نهایی، ارزیابی و پیشنهادات برای تحقیقات آتی ارائه شده است. در بخش ضمائم این مقاله شبکه *LSTM* و ویژگیهای آن توضیح داده شده است.

۲- تحقیقات مرتبط انجام یافته

تحقیقات زیادی در زمینه پیش‌بینی مسیر عابر پیاده از روی تصاویر انجام گرفته است و با گسترش مفاهیم یادگیری عمیق نتایج خیلی خوبی در این زمینه بدست آمده است. با توجه به اینکه در این تحقیق نیز از شبکه *LSTM* استفاده شده است، لذا تحقیقات مرتبط با مفاهیم یادگیری عمیق در این بخش توضیح داده می‌شود. به این منظور تحقیقات انجام گرفته، در دو گروه طبقه بندی شده است.

[1] *Red-Green-Blue-Depth*

[DOI: 10.61882/jgit.13.1.1]

۲-۱- پیش‌بینی مسیر عابر پیاده با استفاده از پردازش تصویر

تکنیک‌های کلاسیک پردازش تصویر توسط محققان برای پیش‌بینی مسیر عابر پیاده مورد استفاده قرار گرفته‌اند .کالینز و رابرت (۲۰۰۳) برای ردیابی مسیر عابر پیاده از روی تصاویر دوبعدی از روش ویژگی پایه میانگین تغییر مکان[1] استفاده کردند. ابعاد پنجره جستجوی اولیه که برای ردیابی انتخاب می‌شود، تاثیر خیلی زیادی در نتایج دارد. همچنین در فرایند ردیابی با توجه به اینکه اندازه عابر پیاده بر روی تصویر با دور و نزدیک شدن به دوربین تغییر می‌کند، لذا ابعاد پنجره جستجو نیز باید متناسب با آن تغییر یابد. ولی در این روش مکانیسم دقیق و سریع برای تعیین ابعاد و مقیاس پنجره جستجو برای ردیابی همزمان وجود ندارد[۲۲]. گرندهی و همکاران (۲۰۰۸) از مجموعه داده اینریا[2] که شامل تصاویر عابر پیاده در مکانهای مختلف می‌باشد، استفاده نموده و تنها با استفاده از شناسایی جهت حرکت عابر پیاده در هر تصویر مسیر آن را پیش‌بینی نمودند. برای این کار ابتدا عابر پیاده شناسایی شده و چارچوب[3]عابر پیاده در تصویر مشخص می‌شود، سپس بردار ویژگیهای عابر پیاده با استفاده از توصیف‌گر هیستوگرام گرادیان جهت‌دار[4] (*HOG*) استخراج شده و در ۸ کلاس (جهت حرکت) با الگوریتم ماشین بردار پشتیبان[5](*SVM*) طبقه‌بندی شده است[۲۳]. سیموسرا و همکاران (۲۰۱۲) از تک‌تصویر برای برآورد موقعیت سه‌بعدی انسان استفاده نمودند. بدین منظور نقاط کلیدی بدن انسان را شناسایی نموده و با فیلترهای *HOG* مقدار تغییرات حرکتی عابر پیاده را بدست آوردند[۲۴]. کوپینترا و همکاران (۲۰۱۴) برای پیش‌بینی مسیر عابر پیاده با استفاده از تصاویر استریو ابر نقاط را تولید کرده و با شناسایی نقاط کلیدی بدن انسان در فضای سه‌بعدی جهت حرکت عابر را تشخیص دادند[۲۵]. کیم و همکاران (۲۰۱۵) برای پیش‌بینی مسیر عابرین پیاده از روی تصاویر، از کالمن فیلتر برای برآورد پارامترهای مدل حرکات انسان بر مبنای موانع سرعت متقابل استفاده نموده‌اند و با ترکیب فیلتر کالمن و برآورد پارامترها در محیط‌های شلوغ، توانسته‌اند برای موارد مشابه استنتاج‌های قابل قبولی ارائه کنند، نام این روش را اصطلاحاً موانع سرعت متقابل بیزی[6] *(Bayesian-RVO)* نامیده‌اند. لازم به ذکر است که روش فوق در محیط‌های چالش بر انگیز مانند صحنه‌های دارای مانع یا قدرت تفکیک پایین نیاز به بررسی و تلفیق با روشهای دیگر دارد[۲۶]. برا و همکاران (۲۰۱۶) با استفاده از اطلاعات مسیر حرکت در چند فریم متوالی و الگوریتم‌های مجموعه کالمن فیلتر[7](*EnKF*) و انتظار حداکثری[8](*EM*) اطلاعات وضعیت عابر پیاده را استخراج کرده و بر مبنای آن جهت حرکت عابرین پیاده را بصورت مجزا و بصورت مجموعه‌ای از عابرین با مدل‌های حرکتی تعیین کرده است. سپس ویژگیهای حرکت کلی و ترکیب شده با الگوهای حرکتی محلی و همچنین با در نظر گرفتن سرعت موانع دینامیک، برای هر عابر، موقعیت آتی را محاسبه نموده است. این روش را که از ترکیب الگوهای حرکتی کلی و محلی استفاده می‌کند اصطلاحاً *GLMP*[9] گویند[۲۷].

۲-۲- پیش‌بینی مسیر عابر پیاده با استفاده از یادگیری عمیق

امروزه از یادگیری عمیق برای حل بسیاری از مسائل

[1] *Mean shift*
[2] *INRIA(Sample Data)*
[3] *Bounding box*
[4] *Histogram of Oriented Gradiant*
[5] *Support Vector Machine*
[6] *Bayesian Reciprocal Velocity Obstacles*
[7] *Ensemble Kalman Filter*
[8] *Expectation Maximization*
[9] *Global and Local Movement Patterns*

[DOI: 10.61882/jgit.13.1.1]

علمی در حوزه های مختلف استفاده میشود و براساس نیاز حوزه‌های کاربردی، الگوریتم‌های متنوعی ارائه شده است. ماهانگ و همکاران (۲۰۱۷) برای پیش‌بینی مسیر حرکت عابرین پیاده پیشنهاد می‌نمایند که مدل ارائه شده باید شامل همه عابرین پیاده باشد، زیرا مسیر حرکت عابرین پیاده تحت تأثیر همدیگر می‌باشد. هر چند رفتار هر شخص می‌تواند متفاوت باشد. بدین منظور از تئوری‌های مورد استفاده در بازی‌های رایانه‌ای[1]کمک گرفته و برای مدل‌سازی رفتار هر فرد با استفاده از تصاویر اخذ شده، از شبکه‌های عصبی عمیق استفاده نمودند[۲۸]. آلاهی و همکاران (۲۰۱۷) الگوریتم شبکه عصبی *LSTM* [۱۶] که برای تشخیص دستخط و صدا مناسب می‌باشد را برای پیش‌بینی مسیر حرکت عابرین توسعه دادند. در این روش، عابرین بر روی تصویر شناسایی شده و برای هر کدام یک شبکه *LSTM* اختصاص داده می‌شود. این شبکه با ردیابی عابر پیاده در تصاویر متوالی وضعیت شخص را یاد گرفته و موقعیت آن را در تصاویر آتی پیش‌بینی می‌کند. شبکه توسعه داده شده را اصطلاحاً شبکه *LSTM* اجتماعی[2] (*S-LSTM*) نامیده‌اند[۲۰]. یو و همکاران (۲۰۱۸) از الگوریتم *LSTM* به دلیل تاثیر رفتار عابرین دیگر بر رفتار انسان، در سه مقیاس مختلف به نام‌های شخص، اطراف شخص و کل تصویر، برای پیش‌بینی مسیر عابر پیاده استفاده نموده و نام آن را[3]*SS-LSTM* می‌نامند. هیو و همکاران (۲۰۱۹) برای تخمین وضعیت عابر پیاده بر مبنای داده‌های آموزشی آنها را در ۸ جهت طبقه‌بندی نموده، سپس برای آموزش شبکه و تولید مدل عمیق از روشی با عنوان معلم-دانش آموز[4] استفاده کرده و حجم داده مورد نیاز برای آموزش را کاهش دادند. در مرحله اول از تکنیک یادگیری عمیق شبکه رزنت[5]استفاده شده است که برای آموزش آن از داده‌های برچسب‌دار یک یا صفر استفاده شده و شامل ۱۰۱ لایه می‌باشد[۲۱]. شی و همکاران (۲۰۱۹) از *S-LSTM* با مقداری تغییرات برای پیش‌بینی مسیر عابر پیاده استفاده نموده و با مطالعه مسیر حرکتی در ۸ فریم(*3.2* ثانیه)، برای ۲۵ فریم بعدی (*10* ثانیه) پیش‌بینی موقعیت را انجام داده‌اند[۷].

همان‌طور که مشاهده می‌شود، در تحقیقات انجام گرفته در زمینه پیش‌بینی وضعیت عابر پیاده با استفاده از شبکه‌های عمیق[۲۹]، از داده‌های *RGB* در فضای دو بعدی استفاده شده است و عامل عمق یا فاصله که در رفتار عابر پیاده تأثیر زیادی دارد، مورد بررسی و مطالعه قرار نگرفته است.

۳- روش پیشنهادی

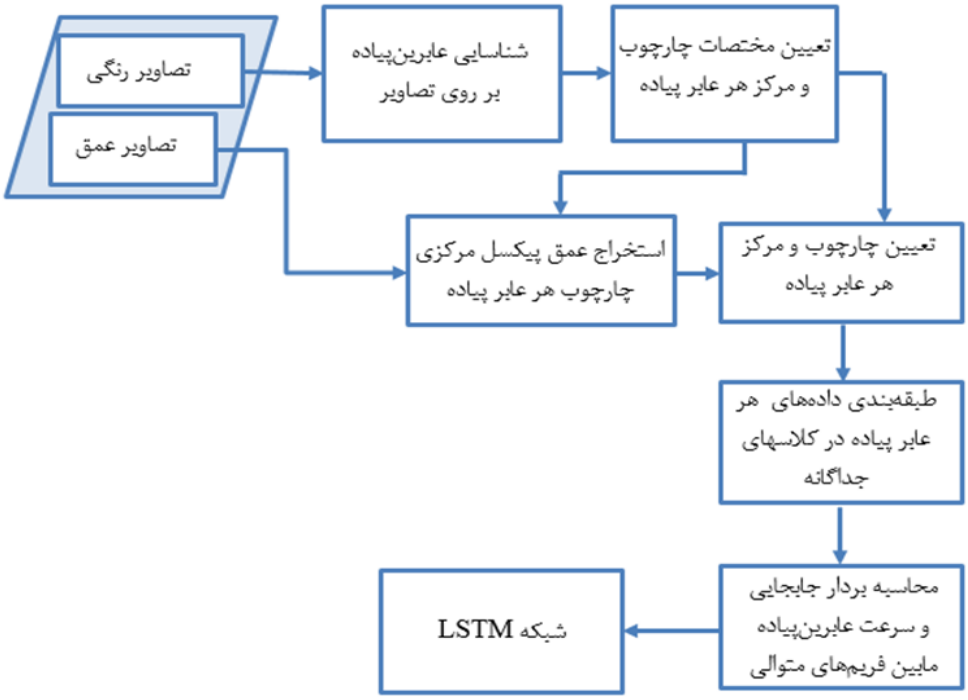
در این روش، برای پیش‌بینی موقعیت عابر پیاده از روی تصاویر، از شبکه یادگیری عمیق *LSTM* استفاده شده است و برای اینکه پیش‌بینی‌ها به دنیای واقعی نزدیک باشد، علاوه بر تصاویر، از اطلاعات عمق نیز استفاده شده و شبکه *3D-LSTM* برای پیش‌بینی موقعیت عابر پیاده پیشنهاد شده است.

داده‌های مورد نیاز شامل تصاویر متوالی از محیط اطراف به همراه داده‌های عمق می‌باشد. عابرین پیاده از روی تصاویر شناسایی شده و پس از تعیین مختصات هر عابر پیاده، اطلاعات عمق مربوطه نیز از روی داده‌های عمق استخراج می‌شود. سپس عابرین در تصاویر متوالی ردیابی شده و مختصات سه‌بعدی هر عابر پیاده بصورت مجزا در کلاس خاص طبقه‌بندی می‌شود. پس از تعیین مقدار جابجایی و سرعت تغییر موقعیت عابرین، داده‌های آماده شده وارد شبکه *LSTM* می‌شوند. این داده‌ها به سه بخش تقسیم می‌شوند. بخش اصلی داده‌ها برای آموزش شبکه استفاده می‌شود. بخش دوم برای ارزیابی آموزش شبکه و بخش سوم

[1] *Game Theory*
[2] *Social LSTM*
[3] *Social Scene LSTM*
[4] *Teacher-Student Framework*
[5] *Res Net*

[DOI: 10.61882/jgit.13.1.1]

برای تست دقت شبکه مورد استفاده قرار می‌گیرد. الگوریتم پیشنهادی در شکل(۲) نشان داده شده است.



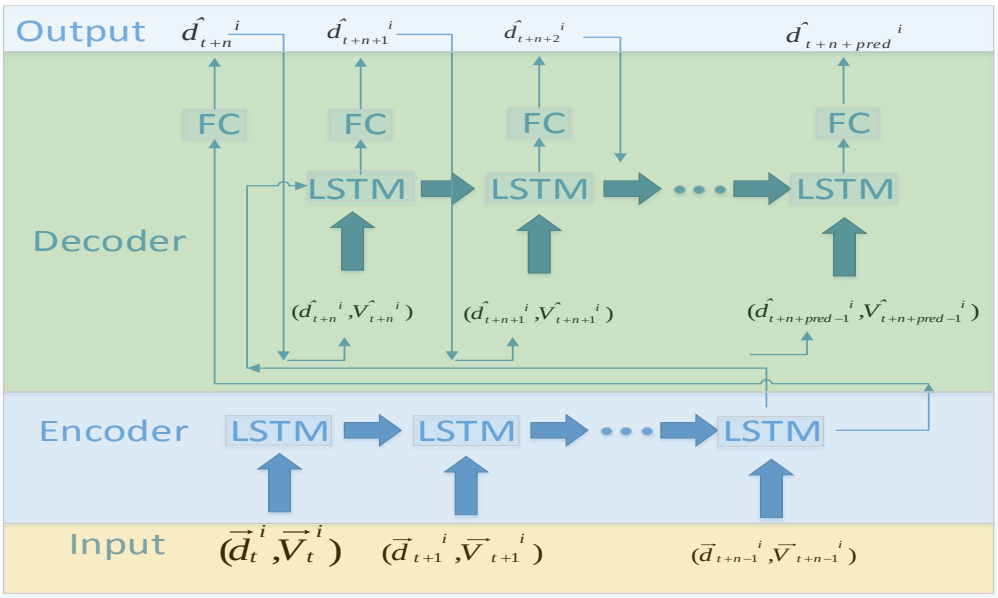
شکل۲: الگوریتم پیشنهادی برای پیش‌بینی موقعیت سه‌بعدی عابر پیاده

۳-۱- معماری شبکه *3D-LSTM* در روش پیشنهادی

ساختار شبکه پیشنهادی در شکل(۳) نشان داده شده است که شامل دو مرحله کدگذاری[1] و کدگشایی[2] است. در مرحله کدگذاری بردار جابجایی و سرعت حرکت عابر پیاده در فریم‌های متوالی از زمان t تا زمان $t+n-1$ به شبکه *LSTM* به عنوان ورودی داده می‌شود و در مرحله کدگشایی مقدار بردار جابجایی و سرعت عابر پیاده در زمان های آتی پیش‌بینی شده و مقدار جابجایی نسبت به فریم قبلی به عنوان خروجی شبکه محاسبه می‌شود. همان‌طور که در ساختار شبکه پیشنهادی نشان داده شده است، مقدار بردار جابجایی و سرعت پیش‌بینی شده در زمان $t+n$ برای پیش‌بینی بردار جابجایی در زمان $t+n+1$ نیز مورد استفاده قرار می‌گیرد. به عبارت دیگر برای پیش‌بینی موقعیت عابر پیاده در فریم‌های آتی (بیشتر از یک فریم) حتی از مقدارجابجایی و سرعت پیش‌بینی شده در فریم‌های قبل از آن نیز استفاده می‌شود. این فرایند برای تمام عابرین پیاده که در تصویر شناسایی می‌شوند، انجام می‌گیرد و برای هر عابر پیاده که در فریم‌های جدید شناسایی می‌شود، بعد از ردیابی در چند فریم متوالی از زمان t_0 تا زمان $t+n-1$، مقدار بردار جابجایی و سرعت طبق روابط (۱) الی (۳) محاسبه شده و به عنوان ورودی وارد شبکه *LSTM* (رابطه(۴)) می‌شوند.

[1] *Encoder*
[2] *Decoder*

[DOI: 10.61882/jgit.13.1.1]



شکل۳: معماری شبکه *3D-LSTM* پیشنهادی

$$\vec{p}_t^{\,i} = (x_t^i, y_t^i, z_t^i) \quad \text{رابطه(۱)}$$

$$\vec{d}_t^{\,i} = (dx_t^i, dy_t^i, dz_t^i) = \overrightarrow{p_t^i} - \overrightarrow{p_{t-1}^i} \quad \text{رابطه(۲)}$$

$$\vec{V}_t^{\,i} = (Vx_t^i, Vy_t^i, Vz_t^i) = \vec{d}_t^{\,i} / dt \quad \text{رابطه(۳)}$$

$$h\_s_{t+1}, c\_s_{t+1} = LSTM(\varnothing_a(\vec{d}_t^{\,i}, \vec{V}_t^{\,i}), h\_s_t, c\_s_t) \quad \text{رابطه(۴)}$$

در رابطه (۱)، $\vec{p}_t^{\,i}$ مختصات سه‌بعدی موقعیت عابر پیاده i ام در زمان t می‌باشد. در رابطه (۲) بردار $\vec{d}_t^{\,i}$ جابجایی موقعیت عابر پیاده i ام مابین دو فریم متوالی در زمان‌های t-1 و t می‌باشد. در رابطه (۳) بردار $\vec{V}_t^{\,i}$ سرعت حرکت عابر پیاده i ام مابین دو فریم متوالی می‌باشد. در رابطه (۴) ورودی‌های شبکه *LSTM* علاوه بر بردار جابجایی و سرعت عابر پیاده i ام، اطلاعات ذخیره شده از وضعیت مراحل قبل در لایه پنهان[1] (h_s_t) و سلول حافظه تا زمان t (C_s_t) می‌باشد.

۳-۲- تابع برآورد خطای شبکه (تابع ضرر)[2]

معمولاً برای تعیین میزان خطا در مرحله آموزش شبکه و تعیین وزن‌های شبکه براساس اختلاف مابین مقدار واقعی و مقدار برآورد شده، از تابعی به نام تابع ضرر استفاده می‌شود. در رابطه (۵) p_t^i موقعیت واقعی عابر پیاده i ام در زمان t و $\hat{p}_t^i$ موقعیت پیش‌بینی شده برای همان عابر پیاده در زمان t می‌باشد. از آنجایی که در این شبکه از یک زیرشبکه[3] همراه با مدل پیش‌بینی متوالی استفاده می‌شود، لذا با تابع انتشار عقبگرد زمانی[4] نمی‌توان به‌طور کامل شبکه را آموزش داد. لذا تابع برآورد خطای شبکه مورد استفاده برای آموزش آن از نوع تابع انتشار عقبگرد کوتاه زمانی[5] می‌باشد. همان‌طور که در شکل(۴) نشان داده شده است، بخش

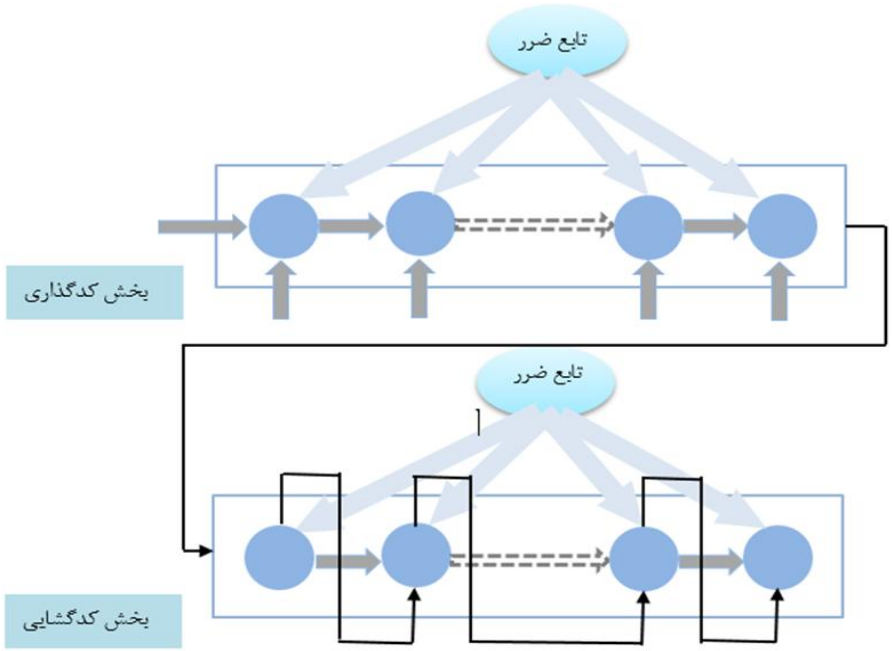
[1] *Hidden State*
[2] *Loss Function*
[3] *Sub Network*
[4] *Back Propagation Through Time*
[5] *Truncated Back Propagation Through Time*

[DOI: 10.61882/jgit.13.1.1]

کدگذاری و کدگشایی به صورت مجزا آموزش داده می‌شود. در بخش کدگذاری مقدار تابع ضرر برای عابر پیاده طبق رابطه(۵) از فریم اول (*T1=1*) الی آخرین فریم که عابر پیاده مشاهده شده است، (*T2=observation*) محاسبه می‌شود و در بخش کدگشایی از *T1=obs+1* الی *T2=observation+prediction* برای محاسبه آن استفاده می‌شود.

رابطه(۵)

$$L(\hat{p}_t^i, p^i{}_t) = \sum_{t=T1}^{T2} \left\| \hat{p}_t^i - p^i{}_t \right\|_2$$



شکل۴: آموزش مدل با تابع انتشار عقبگرد کوتاه زمانی

۴- پیاده‌سازی

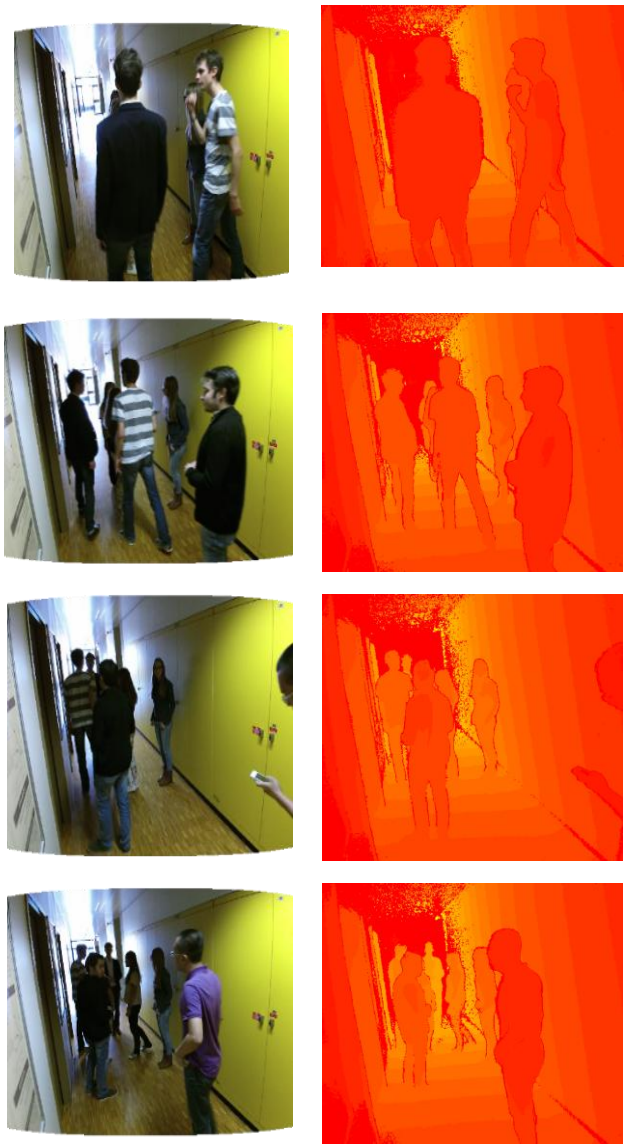
در این بخش ابتدا داده‌های مورد استفاده معرفی می‌شود. سپس نتایج بدست آمده با روش پیشنهادی بررسی و آنالیز می‌شود. نهایتاً نتایج پیش‌بینی با استفاده از داده‌های تست ارزیابی می‌شود.

۴-۱- داده‌های مورد استفاده

اغلب تحقیقات انجام گرفته در زمینه شناسایی و پیش‌بینی وضعیت عابر پیاده از روی تصاویر، بصورت دو بعدی انجام گرفته است، که علت اصلی آن دو بعدی بودن داده‌های تصویری می‌باشد. ولی با توجه به اینکه محیط اطراف عابر پیاده سه‌بعدی است و بعد سوم تاثیر زیادی در رفتار عابر پیاده دارد، لذا در این تحقیق، از داده‌های *RGB-D* تهیه شده در دانشگاه پلی تکنیک لوزان سوئیس[1] *(EPFL)* برای پیش‌بینی موقعیت عابر پیاده استفاده شده است. نمونه‌ای از این تصاویر در شکل (۵) نشان داده شده است. این داده‌ها شامل حدود ۳۰۰۰ فریم می‌باشد که در یک راهرو ساختمان دانشگاه توسط کینکت اخذ شده است.

[1] *University of Lausanne Polytechnic*

[DOI: 10.61882/jgit.13.1.1]



شکل۵: نمونه داده *RGB-D* دانشگاه *EPFL*

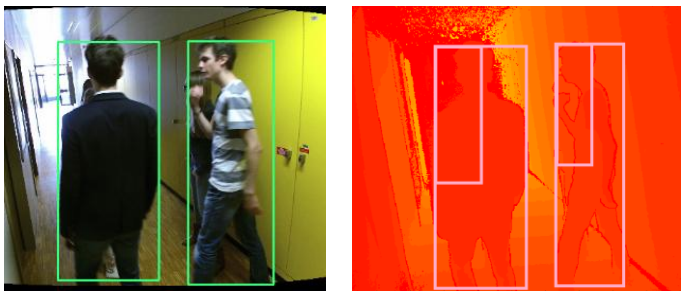
۴-۲- آماده‌سازی داده‌ها

اولین مرحله برای آماده‌سازی داده‌ها، شناسایی عابرین پیاده در تصاویر اخذ شده است. بدین منظور روشهای مختلفی توسط محقّقین توسعه داده شده است. از جمله می‌توان به روشهای *Fast-RCNN* [۳۰] و *Yolo* [۳۱] اشاره نمود. در این تحقیق از شبکه آموزش داده شده و مورد استفاده در تنسور فلو[1] برای شناسایی عابرین پیاده در تصاویر استفاده شده است. برای اجرای روش پیشنهادی، ابتدا کتابخانه‌های مورد نیاز نصب شده و سپس مدل شناسایی عابر پیاده تنسور فلو فراخوانی می‌شود. با توجه به اینکه مدل آموزش داده شده در تنسور فلو بر روی تصاویر خاکستری[2] و با ابعاد

[1] *TensorFlow*
[2] *Grayscale*

[DOI: 10.61882/jgit.13.1.1]

۳۰۰*۳۰۰ پیکسل می‌باشد، لذا در مرحله بعد، تصاویر *RGB-D* فراخوانی شده به ابعاد ۳۰۰*۳۰۰ پیکسل و خاکستری تبدیل شدند. پس از آن شناسایی عابرین پیاده از تصویر انجام گرفت. بعد از شناسایی عابرین، مختصات چارچوب دور عابر پیاده از روی تصاویر استخراج شده و از داده‌های عمق مربوط به همان تصویر فراخوانی شده، مقدار عمق پیکسل مرکزی چارچوب به عنوان مختصات بعد سوم عابر پیاده (شکل(۶)) استخراج شده است. این کار برای تمام عابرین شناسایی شده در تصویر انجام گرفته است. شبه‌کد آماده سازی داده‌ها در الگوریتم (۱) نشان داده شده است.



شکل۶: شناسایی عابرین پیاده و تعیین عمق آن از روی داده‌های *RGB-D*

1. Import necessary libraries
2. Read the frozen graph from a file
3. Load the RGB image and the depth image
4. Resize the RGB image and convert it to grayscale
5. Reshape the resized image to have the format [1, height, width, 3]
6. Extract the image tensor from the session
7. Extract the detection boxes, scores, classes, and num_detection
8. Extract The coordinates of the center of the boxes and its depth

الگوریتم(۱) : شبه کد آماده‌سازی داده‌ها

پس از استخراج موقعیت سه‌بعدی عابرین، داده‌های مربوط به هر عابر در تصاویر متوالی در یک کلاس قرار داده شد. طبق فرایند توضیح داده شده در الگوریتم (۱) از ۳۰۰۰ فریم تصویر پردازش شده، ۱۰۰۰ عابر در تصاویر متوالی ردیابی شدند. در مرحله بعد موقعیت سه‌بعدی عابرین شناسایی شده در تصاویر متوالی در بسته‌های ۲۱ تایی طبقه‌بندی شدند. سپس طبق روابط(۲) و (۳) مقدار بردار جابجایی و سرعت حرکت عابر پیاده مابین فریم‌های متوالی محاسبه گردید. از میان بیش از ۲۰۰۰ بسته ایجاد شده، ۱۲۵۳ بسته برای آموزش شبکه، ۴۱۷ بسته برای ارزیابی[1] و ۴۱۷ بسته برای تست نهایی شبکه در نظر گرفته شد.

۴-۳- پارامترهای شبکه پیشنهادی

ساختار شبکه پیشنهادی برای آموزش و پیش‌بینی موقعیت سه‌بعدی عابر پیاده در شکل (۳) نشان داده شده است. داده‌های آماده شده در مرحله قبل برای آموزش، ارزیابی و تست شبکه که در بسته‌های ۲۰تایی برای هر عابر طبقه‌بندی شده‌اند، به عنوان ورودی

[1] *Evaluation*

[DOI: 10.61882/jgit.13.1.1]

شبکه *LSTM* وارد آن می‌شوند. ابتدا بصورت تصادفی ۱۰۰ بسته از ۱۲۵۳ بسته آموزشی انتخاب شده و شبکه آموزش داده می‌شود و بر مبنای حداقل شدن مقدار تابع هزینه، وزن شبکه برآورد شده و با نرخ آموزش ۰٫۰۰۱ تغییر می‌یابند. برای ارزیابی مدل، از داده‌های ارزیابی که هیچ نقشی در آموزش مدل نداشتند، استفاده می‌شود. در نهایت پس از تکرار این فرایند (۲۵ بار) و تعیین مدل نهایی، از داده‌های تست برای تعیین مقدار خطای مدل پیشنهادی استفاده شده است. پارامترهای پیاده‌سازی شبکه پیشنهادی طبق جدول(۱) می‌باشد. شبه کد آموزش شبکه و تعیین مدل پیش‌بینی در الگوریتم (۲) نشان داده شده است.

```
FUNCTION main(repeat=25):
    SET setting TO dictionary with specific keys and values
    SET keys TO list of specific keys from setting
    SET name TO formatted string using keys and setting values
    CALL prepare_data with setting and UNPACK returned values into train_data, valid_data, test_data

FUNCTION prepare_data(setting):
    Load Data
    SET train, valid, test TO load_data(setting)
    SET train_data TO DataLoader instance with train, batch_size, shuffle flag, and collate_fn
    SET valid_data TO DataLoader instance with valid, batch_size, shuffle flag, and collate_fn
    SET test_data TO DataLoader instance with test, batch_size, shuffle flag, and collate_fn
    RETURN train_data, valid_data, test_data

FUNCTION train_fn(train_data, valid_data, setting):
    SET net TO Predictor instance with setting
    SET optimizer TO Adam optimizer with net parameters and learning rate
    SET loss_fn TO MSELoss
    SET best_loss TO infinity
    FOR I IN range(nepoch):
        SET net to training mode
        SET ep_loss TO 0
        FOR j, batch IN enumerate(train_data):
            SET x, y, m TO batch
            SET yhat TO net(x)
            SET loss TO loss_fn(yhat[m], y[m])
            RESET optimizer gradients
            COMPUTE gradients of the loss
            UPDATE optimizer parameters
            ADD loss to ep_loss
        OUTPUT training RMSE for the current epoch
```

الگوریتم(۲): شبه کد آماده‌سازی داده‌ها

[DOI: 10.61882/jgit.13.1.1]

جدول۱: پارامترهای پیاده‌سازی شبکه *LSTM*

پارامتر	مقدار
تعداد عضو هر بسته[1]	۲۰
نرخ آموزش[2]	۰/۰۰۱
تعداد اپوک[3]	۱۰۰
تعداد لایه	۲
اندازه *Rnn*	۲۰۰
تعداد تکرار	۲۵
تابع بهینه‌سازی[4]	تابع آدام[5]
تابع هزینه[6]	فاصله اقلیدسی حداقل[7]

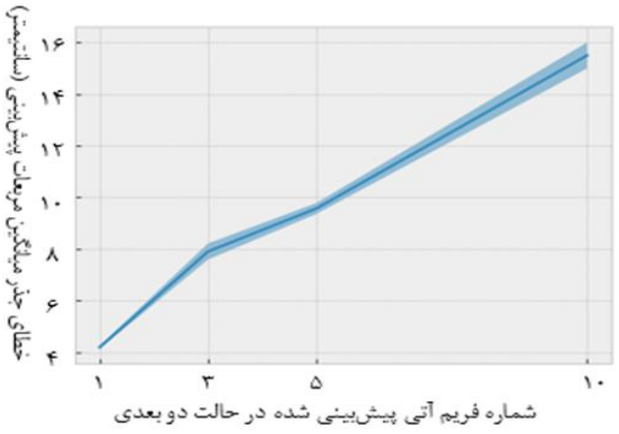
۴-۴- ارزیابی نتایج

برای ارزیابی نتایج پیش‌بینی موقعیت سه‌بعدی عابرین پیاده، مدل بدست‌آمده بر روی داده‌های تست در حالت دو بعدی و سه‌بعدی اجرا گردید. مقدار خطای جذر میانگین مربعات *(RMSE)* پیش‌بینی موقعیت عابر پیاده در حالت سه‌بعدی در مقایسه با حالت دو بعدی نشان می‌دهد که در حالت سه‌بعدی مقدار *RMSE* بدست‌آمده با حالت دوبعدی برابر بوده و تفاوت معنی‌داری مابین آنها وجود ندارد، در حالی که در حالت سه‌بعدی مقدار عمق را نیز پیش‌بینی می‌نماید؛ ولی در حالت دو بعدی هیچ اطلاعاتی از عمق را ارائه نمی‌دهد. با توجه به اینکه محیط پیرامون ما بصورت سه‌بعدی است، پیش‌بینی در حالت سه‌بعدی به واقعیت نزدیکتر بوده و اتخاذ تصمیم مناسب بر مبنای پیش‌بینی موقعیت سه‌بعدی عابر پیاده نسبت به حالت دو بعدی منطقی‌تر می‌باشد.

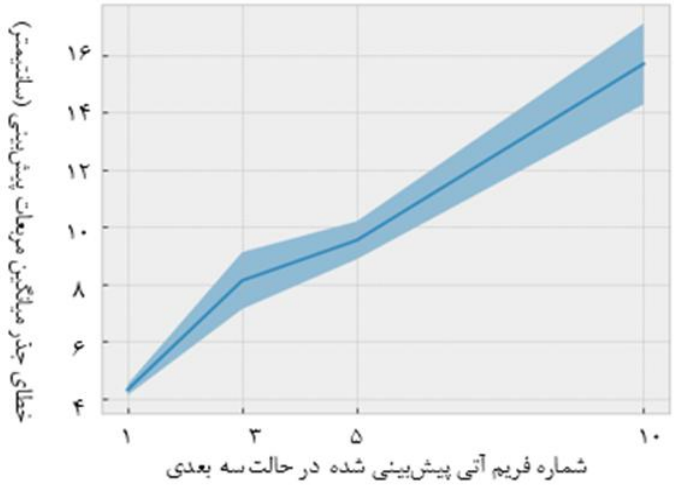
برای مقایسه نتایج در حالت دو بعدی و سه‌بعدی طبق مدل به دست‌آمده پیش‌بینی برای فریم‌های آتی یک، سه، پنج و ده روی داده‌های تست اجرا گردید و مقدار خطای بدست‌آمده طبق شکل(۷) و جدول(۲) می‌باشد.

[1] *Batch Size*
[2] *Learning Rate*
[3] *Epoch*
[4] *Optimization*
[5] *Adam Optimizer*
[6] *Cost Function*
[7] *Min((Euclidean Distance)*

[DOI: 10.61882/jgit.13.1.1]



(الف)



(ب)

شکل۷: نمودارخطای جذر میانگین مربعات پیش‌بینی موقعیت عابر پیاده در فریم‌های آتی، (الف) حالت دو بعدی، (ب) حالت سه‌بعدی

جدول۲: مقدار میانگین خطا و تغییرات آن بر مبنای سانتیمتر

	شماره فریم	۱	۳	۵	۱۰
دو بعدی	میانگین خطای پیش‌بینی	۴/۲	۷/۹۰	۹/۵۷	۱۵/۵۰
	انحراف معیار	۰/۰۱	۰/۰۵	۰/۰۴	۰/۱
سه بعدی	میانگین خطای پیش‌بینی	۴/۳۱	۸/۱۳	۹/۵۴	۱۵/۷۰
	انحراف معیار	۰/۰۳	۰/۱۹	۰/۱۳	۰/۲۸

[DOI: 10.61882/jgit.13.1.1]

همچنین در مقایسه مقدار *RMSE* در این دو حالت ملاحظه می‌شود که مقدار *RMSE* در حالت سه‌بعدی علیرغم اضافه شدن بعد سوم با حالت دو بعدی برابر می‌باشد و مقدار انحراف معیار *RMSE* در هر دو حالت کمتر از پیکسل می‌باشد

و این نشان‌دهنده آن است که مدل بدست آمده برای بسته داده‌های انتخاب شده از مجموعه داده‌های تست، پیش‌بینی‌های مشابهی انجام داده است و مقدار خطای بدست آمده برای فریم‌های آتی مشابه می‌باشد.

۵- نتیجه‌گیری

در این تحقیق بر اهمیت و ضرورت پیش‌بینی موقعیت سه‌بعدی عابران پیاده تاکید شده و بر اساس شبکه *3D-LSTM* یک مدل ارائه شده است. آزمایش‌ها نشان داد که دقت پیش‌بینی موقعیت عابر پیاده در حالت سه‌بعدی علیرغم اضافه شدن بعد سوم برابر با حالت دو بعدی است و با توجه به اینکه پیش‌بینی در حالت سه‌بعدی به دنیای واقعی نزدیکتر است و بعد سوم نقش بالایی در تصمیم‌گیری دارد، پیشنهاد می‌شود پیش‌بینی موقعیت عابر پیاده در فضای سه‌بعدی انجام گیرد. در این مطالعه، پیش‌بینی موقعیت سه‌بعدی عابران پیاده بر روی داده‌های *RGB-D* ارائه شده توسط یک کینکت ثابت از یک راهرو بررسی شده است و برای تحقیقات آتی پیشنهاد می‌شود پیش‌بینی موقعیت سه‌بعدی عابران پیاده با استفاده از داده‌های *RGB-D* تولید شده توسط کینکت نصب‌شده روی عوارض متحرک مانند خودرو انجام گیرد.

۶- ضمائم

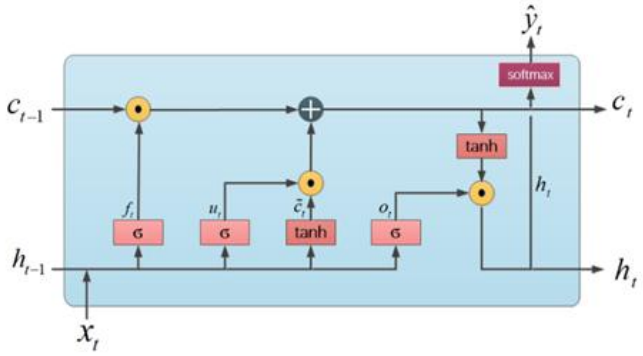
شبکه *LSTM*

شبکه‌های اولیه برای پیش‌بینی خیلی ساده بودند و بر مبنای چندین لایه پنهان و اوزان مربوطه که از طریق آموزش شبکه بدست می‌آمدند خروجی شبکه را محاسبه می‌نمودند. مشکل اساسی این نوع شبکه‌ها عدم توانایی در ذخیره‌سازی اطلاعات نرونها و خروجی‌ها برای پردازش‌های بعدی بود؛ لذا برای رفع این نقص، شبکه‌های بازگشتی مطرح شدند[۱۵].

شبکه *LSTM* در سال ۱۹۹۷ توسط هوکریتر و همکارانش معرفی شد و نوعی از شبکه عصبی بازگشتی است که توانایی ذخیره‌سازی و دسترسی بهتر به اطلاعات برای مدت طولانی را دارد [۱۶]. به عبارت دیگر شبکه *LSTM* برخلاف شبکه عصبی بازگشتی سنتی که در آن محتوا در هر گام زمانی دوباره بازنویسی می‌شود، شبکه *LSTM* قادر است حافظه فعلی را از طریق دروازه‌های ورودی و فراموشی بصورت انتخابی حفظ نموده و یا فراموش کند و بر اساس آنها تصمیم‌گیری نماید. بطور شهودی اگر واحد *LSTM* ویژگی مهمی در دنباله ورودی در گام‌های ابتدایی را تشخیص دهد، می‌تواند این اطلاعات را طی مسیر طولانی منتقل کند[۳۲]. در واقع *LSTM* دارای یک سلول حافظه است که می‌تواند اطلاعات ورودی متوالی را از طریق دروازه‌های ورودی و فراموشی بصورت انتخابی مدیریت نموده و بر مبنای نیاز به هر کدام اجازه ورود و دخالت در پیش‌بینی را بدهد. همچنین قادر است تا قسمتی از شبکه را بصورت تکرار شونده استفاده نماید. معادلات مورد استفاده و ساختار شبکه *LSTM* طبق شکل (۸) است.

در روابط شکل (۸)، x_t بردار ورودی در زمان t ، h_{t-1} و h_t وضعیت پنهان در زمان $t\text{-}1$ و t ، C_{t-1} و C_t سلول حافظه در زمانهای $t\text{-}1$ و t ، W_u, W_f, W_o, W_c ماتریس‌های وزن برای به‌روزرسانی بردارهای u_t, f_t, o_t هستند و b_u, b_f, b_o, b_c بردارهای بایاس می‌باشند.

[DOI: 10.61882/jgit.13.1.1]



$$
\begin{aligned}
u_t &= \sigma(W_u[h_{t-1}, x_t] + b_u), \\
f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f), \\
o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o), \\
\tilde{c}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c), \\
c_t &= u_t \odot \tilde{c}_t + f_t \odot c_{t-1}, \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
$$

شکل۸: شبکه *LSTM* و روابط ریاضی آن

مراجع

[1] *Lee, N., et al. Desire,"Distant future prediction in dynamic scenes with interacting agents", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.*

[2] *Vemula, A., K. Muelling, and J. Oh. Social attention, "Modeling attention in human crowds", in Proceedings of the IEEE international Conference on Robotics and Automation (ICRA), 2018.*

[3] *Combs, T.S., et al., "Automated vehicles and pedestrian safety: exploring the promise and limits of pedestrian detection", American journal of preventive medicine. 56(1): p. 1-7, 2019.*

[4] *Manh, H. and G.J.a.p.a. Alaghband, "Scene-lstm: A model for human trajectory prediction", arXiv preprint arXiv:1808.04018,2018.*

[5] *Rasouli, A. and J.K.J.I.T.o.I.T.S. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice", Proceedings of the IEEE transactions on intelligent transportation systems, 21(3): p. 900-91, 2019.*

[6] *Yazdan, R., M.J.I.J.o.P. Varshosaz, and R. Sensing, "Improving traffic sign recognition results in urban areas by overcoming the impact of scale and rotation", ISPRS Journal of Photogrammetry and Remote Sensing, 171: p. 18-35, 2021.*

[7] *Shi, X., et al., "Pedestrian trajectory prediction in extremely crowded scenarios", Sensors, 19(5): p. 1223,2019*

[8] *Xue, H., D.Q. Huynh, and M. Reynolds, "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction", in Proceeding of the IEEE Winter Conference on Applications of Computer Vision (WACV). 2018.*

[DOI: 10.61882/jgit.13.1.1]

[9] Fernando, T., et al ,.Soft+ hardwired attention, "An lstm framework for human trajectory prediction and abnormal event detection", Neural networks, 108: p. 466-478, 2018

[10] Kalman, R.E., "A new approach to linear filtering and prediction problems", published in Journal of Basic Engineering, 82 (Series D): 35-45. 1960.

[11] Thrun, S., W. Burgard, and D.J.C. Fox ,MA, USA, "Probabilistic Robotics-Intelligent Robotics and Autonomous Agents Series", The MIT Press. 2006.

[12] Williams, C.K., "Prediction with Gaussian processes: From linear regression to linear prediction and beyond, in Learning in graphical models" ,Springer Netherlands. p. 599-621, 1998

[13] Voulodimos, A., et al., "Deep learning for computer vision: A brief review", Computational intelligence and neuroscience, 2018.

[14] Pascanu, R., et al., "How to construct deep recurrent neural networks", 2013.

[15] Hochreiter, S., et al., "Gradient flow in recurrent nets :the difficulty of learning long-term dependencies", A field guide to dynamical recurrent neural networks. IEEE Press,2001

[16] Hochreiter, S. and J.J.N.c. Schmidhuber, "Long short-term memory", Neural computation, p. 1735-1780, 1997

[17] Bahdanau, D., K. Cho, and Y.J.a.p.a. Bengio, "Neural machine translation by jointly learning to align and translate", arXiv preprint arXiv:1409.0473, 2014.

[18] Becker, S., et al., "An evaluation of trajectory prediction approaches and notes on the trajnet benchmark", arXiv preprint arXiv:1805.07663, 2018.

[19] Alahi, A., et al. "Social lstm: Human trajectory prediction in crowded spaces", in Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 961-971, 2016.

[20] Alahi, A., et al.,"Learning to predict human behavior in crowded scenes, in Group and Crowd Behavior for Computer Vision", InGroup and Crowd Behavior for Computer Vision, Academic Press, Elsevier. p. 183-207, 2017

[21] Heo, D., J.Y. Nam, and B.C.J.S. Ko, "Estimation of Pedestrian Pose Orientation Using Soft Target Training Based on Teacher–Student Framework", Sensors, p. 1147, 2019

[22] Collins, R.T. "Mean-shift blob tracking through scale space", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.

[23] Gandhi, T. and M.M. Trivedi. "Image based estimation of pedestrian orientation for improving path prediction", in Proceedings of the IEEE Intelligent Vehicles Symposium, 2008.

[24] Simo-Serra, E., et al. "Single image 3D human pose estimation from noisy observations", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012.

[25] Quintero, R., et al. "Pedestrian path prediction using body language traits", in Proceedings of the IEEE Intelligent Vehicles Symposium Proceedings. 2014.

[26] Kim, S., et al., Brvo: Predicting pedestrian trajectories using velocity-space reasoning. The International Journal of Robotics Research, 34(2), p201-17, 2015.

[27] Bera, A., et al. "GLMP-realtime pedestrian path prediction using global and local movement patterns", in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 2016.

[DOI: 10.61882/jgit.13.1.1]

[28] Ma, W.-C., et al. "Forecasting interactive dynamics of pedestrians with fictitious play", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[29] Ahmadabadian, A.H., et al., "An automatic 3D reconstruction system for texture-less objects", Robotics and Autonomous Systems, 117: p. 29-39, 2019.

[30] Ren, S., et al., "Faster r-cnn: Towards real-time object detection with region proposal networks", IEEE transactions on pattern analysis and machine intelligence, 39(6): p. 1137-1149, 2016

[31] Shafiee, M.J., et al., Fast YOLO, "A fast you only look once system for real-time embedded object detection in video", arXiv preprint arXiv:1709.05943, 2017.

[32] Graves, A.J.a.p.a., "Generating sequences with recurrent neural networks", arXiv preprint arXiv:1308.0850, 2013.

[DOI: 10.61882/jgit.13.1.1]

K. N. TOOSI UNIVERSITY OF TECHNOLOGY
FACULTY OF GEODESY AND GEOMATICS ENGINEERING

Journal of Geospatial Information Technology
Vol.13, No.1, Spring 2025

Research Paper

3D Pedestrian Trajectory Prediction using Deep Learning from Kinect Data

Akbar Jafari [1] , Ali Hosseininaveh [2*] , Mojtaba Mahmoodian[3]

1- Ph.D Student of photogrammetry in department of Geodesy and Geomatics Engineering, K.N.Toosi University of technology
2- Associate Professor of photogrammetry in department of Geodesy and Geomatics Engineering, K.N.Toosi University of technology
3- Associate Professor, Department of Civil and Infrastructure Engineering, Faculty of Engineering, RMIT University, Melbourne, Australia

Abstract

Pedestrian trajectory prediction is a critical challenge in the fields of computer vision and intelligent transportation systems, as it directly impacts the safety and decision-making capabilities of autonomous systems. Most existing approaches rely on two-dimensional (RGB) data and recurrent neural networks such as LSTM (Long Short Term Memory), which neglect the depth dimension and therefore fail to accurately estimate distances between pedestrians and surrounding objects. In this study, we propose a 3D-LSTM (Three Dimension LSTM) model that utilizes RGB-D data obtained from a fixed Kinect sensor to predict pedestrian positions in metric three-dimensional space. The proposed framework includes depth extraction from stereo images, coordinate normalization, and LSTM-based sequence modeling to forecast future pedestrian positions in the (X, Y, Z) coordinates. Experimental evaluations conducted on the École Polytechnique Fédérale de Lausanne (EPFL) dataset demonstrate that the 3D prediction accuracy (average RMSE: 15.7 cm) is comparable to conventional two-dimensional methods while additionally providing real-world distance and spatial interaction information that is crucial for collision avoidance and motion planning. The results indicate that incorporating the third dimension does not degrade performance; instead, it enhances the ability of intelligent systems to make safer and more informed decisions in dynamic environments. This approach lays the groundwork for advanced navigation and autonomous driving systems with enhanced three-dimensional situational awareness.

Key words: *Pedestrian Trajectory, Trajectory prediction, deep learning, 3D-LSTM Network.*

Correspondence Address *: No. 1346, ValiAsr Street, Mirdamad cross, Geomatics Engineering Faculty, K.N. Toosi University of Technology, Tehran, Iran.*
Tel *: +98 21 88882991*
Email*: hosseininaveh@kntu.ac.ir*

[DOI: 10.61882/jgit.13.1.1]