

شبکه عصبی عمیق برای پیش‌بینی تعامل انسان در ویدئو با استفاده از روابط فازی و شار نوری

مه‌لقا افراسیابی^۱، دانشجوی دکتری؛ حسن ختن‌لو^۲، دانشیار؛ محرم منصوری‌زاده^۳، استادیار
^۱ - دانشکده فنی و مهندسی - دانشگاه بوعلی سینا - همدان - ایران - m.afraziabi@basu.ac.ir
^۲ - دانشکده فنی و مهندسی - دانشگاه بوعلی سینا - همدان - ایران - Khotanlou@basu.ac.ir
^۳ - دانشکده فنی و مهندسی - دانشگاه بوعلی سینا - همدان - ایران - mansoorm@basu.ac.ir

چکیده: پیش‌بینی تعامل در ویدئو یکی از موضوعات فعال در بینایی کامپیوتر است، که هدف آن پیش‌بینی تعامل قبل از انجام کامل آن است. این موضوع به دلیل چالش‌های موجود در این زمینه هنوز مورد توجه است. در این مقاله یک شبکه عصبی عمیق برای پیش‌بینی تعامل با استفاده از روابط فازی و شار نوری ارائه شده است. نوآوری این روش ایجاد دو تصویر فازی از یک ویدئو است. این تصاویر فازی بر مبنای گرادیان و شار نوری ایجاد می‌شود. توابع عضویت فازی مناسب برای روابط مکانی بین افراد در حال تعامل در تصاویر گرادیان و شار نوری ایجاد شده است. از طرفی یک تابع عضویت فاصله برای ارزش‌دهی به فریم‌ها و یک تابع عضویت فاصله برای ارزش‌دهی به ناحیه‌ی بین افراد در حال تعامل تعریف شده است. سپس ویژگی‌های مناسب مکانی-زمانی از این تصاویر با استفاده از معماری شبکه عصبی کانولوشن استخراج شده است. نتایج این روش بر روی دو مجموعه داده استاندارد تشخیص تعامل، BIT و UT ارزیابی شده است. نتایج نشان می‌دهد ایجاد تصاویر فازی و استخراج ویژگی‌های عمیق از آن تصاویر باعث افزایش دقت پیش‌بینی تعامل نسبت به روش‌های پیشین شده است.

واژه‌های کلیدی: رابطه مکانی فازی، گرادیان، شار نوری، شبکه کانولوشن.

Deep neural network for interaction prediction in video using fuzzy relationship and optical flow

Mahleqa Afrasiabi¹, PhD student; Hassan Khotanlou², Associate professor; Moharram Mansoorizadeh³, Assistant professor

1- Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran, Email: m.afraziabi@basu.ac.ir

2- Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran, Email: Khotanlou@basu.ac.ir

3- Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran, Email: Mansoorm@basu.ac.ir

Abstract: The aim of interaction prediction in videos is to predict the interaction before it actually happens. Recently, this task has been important in computer vision domain and is gaining a lot of attention due to its challenges. In this paper, a deep neural network using fuzzy relationship and optical flow is proposed to deal with the problem. In this approach for each frame of a given video, first, two fuzzy images are obtained based on the gradient and the optical flow of the frame. Then, two set of features are extracted by a convolutional neural network trained on these images. Final prediction is made by aggregating the two outputs of the network. The proposed method shows promising results on two interaction datasets, namely BIT-Interaction and UT-Interaction.

Keywords: Fuzzy spatial relationship, gradient, optical flow, convolutional neural network.

تاریخ ارسال مقاله: ۱۳۹۷/۱۰/۱۰

تاریخ اصلاح مقاله: ۱۳۹۷/۱۲/۲۰، ۱۳۹۸/۰۲/۲۴ و ۱۳۹۸/۰۳/۱۸

تاریخ پذیرش مقاله: ۱۳۹۸/۰۴/۰۴

نام نویسنده مسئول: حسن ختن‌لو

نشانی نویسنده مسئول: دانشکده فنی و مهندسی - دانشگاه بوعلی سینا - همدان - ایران.

۱- مقدمه

تشخیص و پیش‌بینی کنش و فعالیت یکی از مباحث مهم در بینایی کامپیوتر است. در تشخیص کنش، به بررسی رفتار یک انسان و در تشخیص فعالیت به بررسی رفتار بین دو یا چند فرد (مانند دست‌دادن) یا تعامل انسان و یک شی (مانند بستن پنجره) پرداخته می‌شود. تشخیص فعالیت به دلیل پیچیدگی رفتار افراد در زمان تعامل با یکدیگر و واکنش‌های متفاوت آنان، نسبت به تشخیص کنش دشوارتر است. در پیش‌بینی، هدف تشخیص فعالیت قبل از اتمام آن خواهد بود [۱]. به‌طور مثال همان‌طور که در شکل ۱ نشان داده شده، هنگامی که قسمتی از ویدئو دیده شده انتظار می‌رود بتوان نوع فعالیت در قسمت دیده‌نشده ویدئو را پیش‌بینی کرد. هدف این مقاله پیش‌بینی فعالیت بین دو فرد است. از کاربردهای پیش‌بینی تعامل و فعالیت می‌توان به شناسایی حوادث خطرناک در محیط‌های نظارتی و شناسایی تعامل انسان و ربات نام برد.

از چالش‌های موجود در زمینه پیش‌بینی تعامل می‌توان به اختلاف ظاهر و سرعت حرکت افراد اشاره کرد. از طرفی چالش‌های محیط مانند تغییرات نور، پیش زمینه شلوغ و تصادم بین فرد مورد بررسی و محیط، بر روی روش‌های تشخیص تعامل تأثیرگذار خواهند بود [۲، ۳]. برای پیش‌بینی تعامل نیاز به فهمیدن رفتار افراد در طول زمان است. بنابراین تشخیص تعامل بر اساس یک فریم کاری دشوار است، روش‌های زیادی در این زمینه ارائه شده‌اند که از ویژگی حرکتی مانند انرژی حرکت تصویر^۱ (MEI) و تصویر تاریخیچه حرکت^۲ (MHI) استفاده کرده‌اند. امروزه شبکه‌های عصبی عمیق نشان داده‌اند که توانایی بسیار بالایی برای استخراج ویژگی دارند [۴]. از بین این شبکه‌ها، شبکه‌های عصبی کانولوشن (CNN)^۳ [۵] برای تشخیص کنش و شناسایی اشیاء به کار رفته‌اند، و نتایج بسیار خوبی داشته‌اند. برای استخراج اطلاعات حرکتی از یک مجموعه فریم پی‌درپی شار نوری^۴ استفاده می‌شود. این فریم‌ها به شبکه CNN داده می‌شوند تا اطلاعاتی زمانی- مکانی را استخراج کند. علاوه بر آن ویژگی‌های CNN که از فریم‌ها استخراج شده به روش‌های دسته‌بندی مانند شبکه‌های LSTM^۵ داده می‌شوند تا با استفاده از پردازش توالی ویژگی به پیش‌بینی تعامل پرداخته شود.

در سال‌های گذشته روش‌های فازی مانند سیستم استنتاج فازی

فعالیت با مدل کردن غیرقطعی ویژگی‌ها نتایج خوبی داشته‌اند [۶-۹]. این روش‌ها توانایی مدل‌سازی و پیش‌بینی فعالیت‌های پیچیده انسان را دارند. از آنجایی که انجام هر کنش یا تعامل ماهیتی فازی دارد و نمی‌توان حرکت مشخصی برای تعریف آن در نظر گرفت، استفاده از روش فازی می‌تواند باعث افزایش دقت تشخیص فعالیت شود.

در این مقاله روشی جدید برای نمایش اطلاعات حرکتی برای پیش‌بینی تعامل ارائه شده است. در این روش برای نمایش اطلاعات حرکتی یک تصویر فازی ایجاد می‌شود. این تصویر حاوی اطلاعات مهم حرکت افراد شرکت‌کننده در تعامل است. تصویر فازی ارائه شده بر محدودیت تصاویر رنگی که نیاز به کنتراست بالا برای جدا کردن افراد شرکت‌کننده در تعامل و پیش‌زمینه است، غلبه می‌کند. همچنین تصویر ایجاد شده ترکیب وزن‌داری از اطلاعات حرکتی مناسب در هر ناحیه از مراحل انجام تعامل است.

ورودی روش ارائه شده تصاویر ویدئوهای RGB است. ابتدا دو تصویر فازی از فریم‌های ویدئو ایجاد می‌شود که این تصاویر بر مبنای رابطه فازی بین افراد در حال تعامل است. هر تصویر به یک شبکه CNN، داده می‌شود تا ویژگی‌های مناسب از آن استخراج شود، در نهایت از شبکه تماماً متصل برای پیش‌بینی تعامل استفاده می‌شود. ویژگی این روش این است، که نیازی به جدا کردن پیش‌زمینه نیست و تصویر فازی استخراج شده دارای اطلاعات حرکتی برای پیش‌بینی نوع تعامل است. روش ارائه شده بر روی دو مجموعه داده تعامل انسان UT و BIT تست شده است. نتایج نشان می‌دهد دقت بالاتری نسبت به روش‌های پیشین به دست آمده است، به‌طور مثال دقت تشخیص بر روی مجموعه BIT حدود ۴ درصد افزایش یافته است.

نوآوری روش ارائه شده در ادامه بیان شده است:

- ارائه یک تصویر فازی بر مبنای گرادبان تصاویر برای پیش‌بینی تعامل
 - ارائه یک تصویر فازی بر مبنای تصاویر شار نوری برای پیش‌بینی تعامل
 - ارائه یک شبکه عمیق برای پیش‌بینی تعامل در ویدئو با استفاده از روابط فازی و شار نوری.
- ادامه این مقاله به صورت زیر ساختار بندی شده است. در بخش ۲، کارهای مرتبط مورد بحث قرار گرفته، در بخش ۳، پیش‌زمینه مطالب



شکل ۱: پیش‌بینی تعامل انسان در ویدئو، هدف پیش‌بینی تعامل این است که قبل از اینکه تعامل کامل اتفاق بیافتد نوع تعامل تشخیص داده شود.

مرتب اول، مدل مخفی مارکوف فازی و روش‌های هیبریدی در تشخیص مرتبط به مقاله بیان شده و در بخش ۴ روش پیشنهادی تشریح شده

آید. با توجه به حرکت نقاط اسکلت یک سیستم تشخیص بر مبنای الگوریتم HMM ارائه شده، که Fuzzy HMM نام دارد. در این روش فاصله هر بردار مشاهده بر مرکز دسته‌بند محاسبه می‌شود و عکس این فاصله درجه عضویت بردار مشاهده به دسته‌بند است.

دومین دسته ویژگی‌ها، ویژگی‌های سطح میانه هستند. معمولاً این ویژگی‌ها بر پایه ویژگی‌های سطح پایین هستند. Poselet جز این دسته ویژگی‌هاست، که حالت انسان‌ها را تحت یک دید معین توصیف می‌کند. در روشی از Poselet مبتنی بر HOG با Poselet های مدل بسته کلمات که حاوی توصیف‌گرهای انبوه (SIFT, HOG, MBH) است، برای تشخیص فعالیت استفاده شده است [۲۰].

به‌منظور یادگیری حرکات پیچیده، در روشی حرکات انسان با ترکیبی از کوچک‌ترین واحد سازنده الگوهای حرکتی انسان به نام dyneme نشان داده شده است [۲۱]. Dyneme کوچک‌ترین واحد حرکت انسان است و حرکت کامل بدن یک توالی از حرکت این واحدهای کوچک می‌باشد. بردار فازی تدریجی (FVQ) به‌عنوان تابعی است که گذار بین تصمیم‌گیری‌های نرم را کنترل می‌کند و تصمیم‌گیری‌های نرم برای ترسیم یک بردار موقعیت ورودی در فضای dyneme به‌کاررفته شده است. در نهایت هر حرکت به‌صورت یک مدل حرکت فازی با محاسبه میانگین حساسی متشکل از حالت‌های مختلف حرکت در فضای dyneme نمایش داده شده است.

این دو دسته ویژگی سطح پایین و سطح میانه، ویژگی‌های دستی^{۱۷} هستند که طراح خبره بر اساس تأثیر آن بر دقت و کارایی روش، آن را انتخاب می‌کند [۲۲]. دسته سوم استخراج ویژگی، ویژگی‌های بدون ناظر است. ویژگی‌هایی که بر مبنای ویژگی‌های دستی است، قدرت کافی برای استخراج اطلاعات حرکتی مهم به‌خاطر از دست دادن ساختار تصویر ندارند. اخیراً شبکه‌های عصبی عمیق مانند CNN نشان داده‌اند، که نتایج خوبی نسبت به روش‌های سنتی دارند. به این ویژگی‌ها، ویژگی‌های عمیق گویند که مستقل از نوع دسته‌بند خاص است و مستقیم از تصاویر به‌دست می‌آید [۲۲]. این ویژگی‌ها مقاومت بیشتری به تغییر پذیری درون کلاسی دارند، در نتیجه قدرت تعمیم بالایی دارند [۲۲].

در مقالاتی که هدف تشخیص تعامل یا کنش در ویدئوهای RGB-D است، از روش‌های فازی برای تشخیص استفاده شده است. به‌طور مثال یک نمایش جدید بر مبنای روش فازی ارائه شده [۲۳]، که در این روش ابتدا یک الگوی حرکتی بر مبنای MHI و MEI از فریم‌ها ایجاد می‌شود، سپس بر اساس رابطه زمانی فازی فریم‌ها، الگوها با یکدیگر ترکیب می‌شوند. خروجی این تصویر به یک شبکه CNN داده شده است. این روش برای تشخیص کنش مناسب است. در مقاله [۲۴] برای تشخیص کنش ابتدا یک پیش‌پردازش با استفاده از فیلتر گابور برای جدا کردن پس‌زمینه استفاده شده، سپس با CNN ویژگی استخراج و در نهایت وزن‌دهی فازی به ویژگی انجام شده است. در مقاله [۲۵] نیز برای تشخیص شی از وزن‌دهی فازی بعد از استخراج ویژگی CNN از تصاویر

است. بخش ۵ نتایج تجربی و تحلیل را ارائه می‌کند. در نهایت، در بخش ۶ نتیجه‌گیری بیان شده است.

۲- کارهای پیشین

کارهای مرتبط در زمینه پیش‌بینی تعامل و فعالیت بر اساس انتخاب ویژگی به سه دسته تقسیم می‌شوند [۱۰]:

در اولین گروه، ویژگی‌های سطح پایین استخراج می‌شود. این نوع ویژگی‌ها اطلاعات مکانی و زمانی را استخراج می‌کنند. ویژگی‌های هیستوگرام گرادینان^{۱۸} (HOG)، هیستوگرام شار نوری^{۱۹} (HOF)، هیستوگرام مرز حرکت^{۲۰} (MBH) و نقاط موردنظر زمان-مکان^{۲۱} (STIP) جز این دسته ویژگی‌ها هستند، که در تشخیص تعامل استفاده شده‌اند.

در مقاله [۱۱] از ترکیب ویژگی‌های HOG، HOF و MBH ویژگی HM^{۱۱} ارائه شده، که از الگوریتم SVM^{۱۱} برای تشخیص تعامل استفاده شده است.

در [۱۲] ویژگی‌های خط‌سیر چگال، HOG، HOF، MBH، STIP و تبدیل ویژگی نامتغیر مقیاس^{۱۲} (SIFT^{۱۲}) از یک ویدئو استخراج شده، سپس هر بردار ویژگی با استفاده از تکنیک فیشر به یک بردار به طول ثابت تبدیل شده، سرانجام با استفاده از الگوریتم SVM نوع تعامل تشخیص داده شده است.

در روش MTSSVM^{۱۳} هر ویدئو را به تعداد بخش‌های مساوی تقسیم، سپس نقاط موردنظر و خط‌سیر چگال از هر بخش استخراج، و از مدل بسته کلمات^{۱۴} برای نمایش هر بخش استفاده شده است. برای تشخیص تعامل الگوریتم MTSSVM به‌کاررفته، که یک مدل SVM ساختاریافته است که محدودیت تکاملی حرکت زمانی و محدودیت سازگاری برچسب به آن اضافه شده است [۱۳]. در روش MMAPM^{۱۵} [۱۴] که بهبود یافته MTSSVM است از ترکیب کرنل‌های غیرخطی در الگوریتم SVM برای تشخیص و پیش‌بینی تعامل استفاده شده است.

در [۱۵] ویژگی‌ها را با استفاده از یک توصیف‌گر مکانی-زمانی، محاسبه کرده‌اند، سپس روابط زمانی و علیت را با استفاده از حوزه‌های تصادفی مارکوف مدل‌سازی کرده‌اند. یک روش دیگر بر مبنای یادگیری گراف‌های CRF برای مدل‌سازی تعامل‌های انسانی فراهم شده [۱۶]، که در آن HOG و HOF از بخش‌های تصویر بدن انسان استخراج شده است.

سیستم استنتاج فازی توانایی مدل کردن عدم قطعیت و ترکیب ویژگی‌های مختلف را دارد. از این سیستم با ورودی ویژگی‌های سطح پایین برای تشخیص فعالیت استفاده شده است [۱۷، ۱۸]. در مقاله [۱۸] که هدف آن تشخیص فعالیت است، ویژگی زمانی-مکانی استخراج شده است. ابتدا تصویر شی^{۱۶} و سپس ویژگی‌هایی از آن که شامل اطلاعات ظاهری و سرعت است، استخراج می‌شود. این ویژگی‌ها به‌عنوان ورودی به سیستم استنتاج فازی داده می‌شود. این کار برای یادگیری توابع عضویت از الگوریتم FCM استفاده کرده است.

روش HMM در تشخیص فعالیت بسیار استفاده شده است. در [۱۹] ابتدا تصویر شی استخراج شده، سپس اسکلت تصویر به‌دست می

این تصویر فازی برای حفظ اطلاعات حرکتی استفاده می‌شود و استفاده از CNN باعث استخراج ویژگی محلی از این تصویر می‌شود. برتری این روش سرعت بالا [۲۹] و عدم وابستگی به مقیاس و انتقال و حفظ اطلاعات حرکتی دانست [۲۲]. چون از کل فریم‌ها یک تصویر ایجاد شده و برای استخراج ویژگی به CNN ارسال می‌شود، می‌توان گفت سرعت آن از روش‌هایی که نیاز به ارسال هر فریم به‌تنهایی به CNN است، بالاتر خواهد بود. این روش نسبت به روش‌هایی که بر مبنای LSTM با بار محاسباتی بالایی هستند نیز سریع‌تر خواهد بود. استفاده از ویژگی‌های عمیق نیز باعث افزایش قدرت تعمیم روش شده است.

۳- پیش‌زمینه

از آنجایی که در روش ما از رابطه مکانی فازی، اطلاعات گرادیان، تصاویر کد شده شار و CNN استفاده شده، در این قسمت مختصراً این موارد توضیح داده شده‌اند.

۳-۱- رابطه مکانی فازی

مشخص کردن رابطه مکانی بین اشیا و انسان‌ها در یک تصویر در بینایی کامپیوتر امر مهمی است [۳۰]. در سال‌های گذشته توجه زیادی به تحلیل روابط فازی مکانی شده است [۳۱]. که در حوزه‌های گوناگون مانند تحلیل حرکت انسان و ربات [۳۲] و تحلیل تصاویر پزشکی [۳۳]، به‌کاررفته است. توانایی تمایز بین رابطه‌های مکانی شبیه‌به‌هم، به‌خصوص در فضای ویژگی‌های دوبعدی کار دشواری است. استفاده از متغیرهای زبانی فازی رویکرد خوبی در فرآیند استدلال روابط مکانی فازی دارد.

تعدادی از تحقیقات در زمینه تعریف رابطه مکانی فازی است [۳۳]. در این مقالات، رابطه مکانی شامل راست، چپ، نزدیک و دور است. برای مدل کردن این روابط نیاز به تعریف متغیرهای زبانی فازی است. با توجه به اهمیت فاصله در پردازش تصویر، بیان فازی آن برای حل مسائل این حوزه بسیار مفید است. برای مثال "A نزدیک B است"، یک مدل فازی نزدیک، یک مقدار عددی μ_{near} به جفت (A,B) نسبت می‌دهد که نشان‌دهنده درستی جمله بالا است. تعاریف گوناگونی برای بیان نزدیکی در روابط مکانی فازی تعریف شده که به مسئله موردبررسی بستگی دارد.

رابطه فازی چپ و راست بر اساس توابع $\sin^2 \theta$ یا $\cos^2 \theta$ عضویت مثلثی، دوزنقه‌ای و توابع دیگر قابل تعریف است. این توابع عضویت با μ_{right} و μ_{left} نشان داده می‌شوند، که این مقدار نشان‌دهنده رابطه جهت‌دار بین دو شیء است. معمولاً در هر شیء یک نقطه مرجع در نظر گرفته می‌شود، سپس مقدار عضویت بر اساس درجه رضایت رابطه مکانی برای نقاط اطراف آن نقطه مرجع تعریف می‌شود. به‌طور مثال در شکل ۲ یک شیء مرجع و رابطه مکانی فازی چپ آن نشان داده شده است.

استفاده شده است. برخلاف روش ارائه شده که ابتدا تصویر فازی ایجاد شده، سپس از CNN استفاده می‌شود، در این دو روش [۲۴، ۲۵] ابتدا ویژگی‌های CNN استخراج می‌شوند، سپس از وزن‌دهی فازی در مرحله دسته‌بندی استفاده می‌شود.

در [۲۶] روشی ارائه داده‌اند که در آن تصاویر کدگذاری شده شار، از فریم‌های متوالی محاسبه شده و به شبکه CNN برای استخراج ویژگی‌های عمیق زمانی ارسال شده‌اند. در کار دیگر که گسترش یافته این روش است از گرفتن اطلاعات سراسری و محلی استفاده شده است [۲۷]. این روش شامل چهار مدل اصلی است. در مدل‌های مکانی و فضایی، تصاویر رنگی و تصاویر شار نوری، و در مدل‌های ساختاری مکانی و زمانی، اطلاعات شار نوری و رنگی تصویر هر فرد در حال تعامل و قسمت بالا و پایین بدن آن به شبکه‌های متمایز CNN ارسال شده است. ویژگی عمیق مختلف استخراج شده به شبکه‌های حافظه بلندمدت (LSTM) منتقل شده و سپس از آن یک لایه به‌طور کامل متصل شده‌اند. چهار امتیاز حاصل از مدل مکانی، مدل زمانی، ساختاری مکانی و ساختار زمانی با یکدیگر ترکیب شده تا نوع تعامل پیش‌بینی شود.

در مقاله [۲۸] یک مدل تکراری بلندمدت (LRCN) پیشنهاد شده، که یک شبکه CNN همراه با LSTM است. آن‌ها همچنین هر دو ورودی‌های RGB و شار را در نظر می‌گیرند. برای هر فریم یک بردار ویژگی عمیق استخراج شده و به یک LSTM ارسال شده است، سپس خروجی‌های لایه softmax شبکه در تمام فریم‌ها به‌طور میانگین گرفته شده و محتمل‌ترین برچسب یافته شده است. در این روش‌ها ویژگی‌های عمیق استخراج می‌شوند، روش‌های طبقه‌بندی مانند LSTM مورد استفاده قرار گرفته‌اند، اما معایب این روش را می‌توان با فرآیند محاسبات بالا بیان کرد.

هر دو دسته ویژگی سطح پایین و سطح میانه نسبت به ویژگی بدون ناظر دقت کمتری دارند، از طرفی مشکل روش‌های فازی ارائه شده در مقاله‌های [۱۷، ۱۸] جدا کردن تصویر انسان از زمینه است. که دقت روش وابسته به دقت استخراج تصویر شیء می‌باشد. در ویدئو که نیاز به بررسی ویژگی‌ها در طول زمان است، پیدا کردن ویژگی مناسب که بیانگر اطلاعات ویدئو باشد کار ساده‌ای نیست. امروزه با قدرت ماشین‌های موازی (مانند GPU و CPU clusters) شبکه‌های کانولوشن عصبی توسعه پیدا کرده‌اند، که در تشخیص کنش در ویدئو استفاده شده‌اند. اما ویژگی‌های استخراج شده با استفاده از این شبکه‌ها بصری هستند و به‌تنهایی برای پیش‌بینی تعامل محدودیت دارند. به همین دلیل در این مقاله قبل از استخراج ویژگی‌های عمیق تصویر فازی بر مبنای حرکت افراد ایجاد می‌شود تا قدرت استنتاج را بالا ببرد. این روش شامل دو مرحله است. ابتدا با استفاده از ویژگی‌های سطح پایین یک تصویر فازی از کل ویدئو استخراج می‌شود. در مرحله بعد از یک CNN برای استخراج ویژگی از این تصویر استفاده شده است.

$$I(x, y, t) = I(x + v_x, x + v_y, t + 1) \quad (۳)$$

برای هر بردار شار نوری $V = [v_x, v_y]^T$ هر پیکسل، اندازه (ρ) و جهت $(-\pi \leq \theta \leq \pi)$ از رابطه (۴) محاسبه می‌شود [۳۷].

$$\rho = \sqrt{v_x^2 + v_y^2} \quad (۴)$$

$$\theta = \arctan\left(\frac{v_y}{v_x}\right)$$

۴-۳- شبکه‌های عصبی کانولوشن

شبکه‌های عصبی کانولوشن (CNN) یکی از شبکه‌های عمیق می‌باشند، که نتایج خوبی برای دسته‌بندی تصاویر دارد. این شبکه دارای چند لایه و یک لایه تماماً متصل است. لایه کانولوشن روی تمام تصویر با کرنل‌های با محدوده کوچک را کانوالو می‌کند. این لایه‌ها باعث استخراج ویژگی‌های مناسب از تصاویر هستند. هر لایه کانولوشن یک لایه ادغام ماکزیمم دارد. این لایه باعث کاهش دامنه ویژگی‌ها می‌شود. سرانجام لایه تماماً متصل قرار دارد که پردازش نهایی را انجام می‌دهد، و منجر به کلاس‌بندی می‌شود [۵].

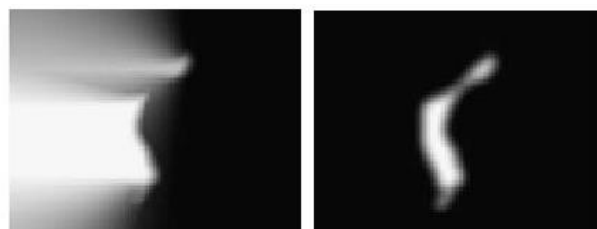
۴-۲- روش پیشنهادی

همان‌طور که در فلوچارت شکل ۳ نشان داده شده، در این مقاله روشی برای پیش‌بینی تعامل که دارای سه بخش اصلی است، ارائه شده است. در بخش اول یک تصویر فازی بر مبنای گرادیان و در بخش دوم یک تصویر فازی بر مبنای شار نوری از تمام فریم‌های ویدئو ایجاد خواهد شد. تصویر فازی گرادیان از کل فریم‌ها ایجاد می‌شود تا اطلاعات حرکتی بین فریم‌ها را نشان دهد و تصویر شار نوری بر مبنای جهت تغییرات حرکت افراد است. سپس هر تصویر به یک شبکه عصبی عمیق داده خواهد شد، تا ویژگی‌های مناسب از هر تصویر استخراج شود. در لایه آخر این شبکه عمیق یک لایه softmax قرار داده شده که احتمال هر تصویر به کلاس‌های تعامل محاسبه می‌شود. نتیجه خروجی این دو شبکه با یکدیگر ترکیب شده تا نوع تعامل پیش‌بینی شود. در ادامه هر بخش شرح داده خواهد شد.

۴-۱- ایجاد تصویر گرادیان فازی

همان‌طور که در شکل ۴ نشان داده شده، ابتدا جعبه‌کران دو فرد در حال تعامل با استفاده از الگوریتم ردیابی [۳۸] استخراج شده، سپس روی هر فریم یک فیلتر میانگین اعمال شده تا تصویر هموار شود، در مرحله بعد (شکل ۴ الف)) گرادیان هر تصویر محاسبه شده است.

سپس مرکز جعبه کران‌ها در نظر گرفته شده است. پیکسل‌هایی که بین دو فرد هستند اهمیت بیشتری خواهند داشت و ناحیه بین دو مرکز اهمیت زیادی در تشخیص خواهد داشت و هر چه از ناحیه مرکز جعبه کران در جهت مخالف حرکت دو فرد دور شده، اهمیت آن نواحی کمتر خواهد شد. برای هر تعاملی نمی‌توان رابطه مکانی فاصله را به‌طور



شکل ۲: الف) شی مرجع، ب) رابطه مکانی چپ برای شی مرجع [۳۳]

۳-۲- اطلاعات گرادیان

ویژگی گرادیان مکانی برای استخراج ویژگی‌های سطح پایین تصویر استفاده می‌شود. این ویژگی قابلیت استخراج لبه‌ها در تصویر را دارد. ابتدا تصویر با استفاده از کرنل $[1 \ 0 \ -1]$ در جهت x و y فیلتر می‌شود تا گرادیان تصویر در جهت x و y محاسبه شود.

برای محاسبه گرادیان هر پیکسل (x, y) در راستای محور افقی و عمودی به ترتیب $g_x(x, y)$ و $g_y(x, y)$ از رابطه (۱) محاسبه می‌شوند.

$$g_x(x, y) = -f(x, y - 1) + f(x, y + 1) \quad (۱)$$

$$g_y(x, y) = -f(x - 1, y) + f(x + 1, y)$$

که $f(x, y)$ شدت پیکسل (x, y) در تصویر است. اندازه بردار گرادیان از رابطه (۲) به دست می‌آید.

$$|g(x, y)| = \sqrt{g_x(x, y)^2 + g_y(x, y)^2} \quad (۲)$$

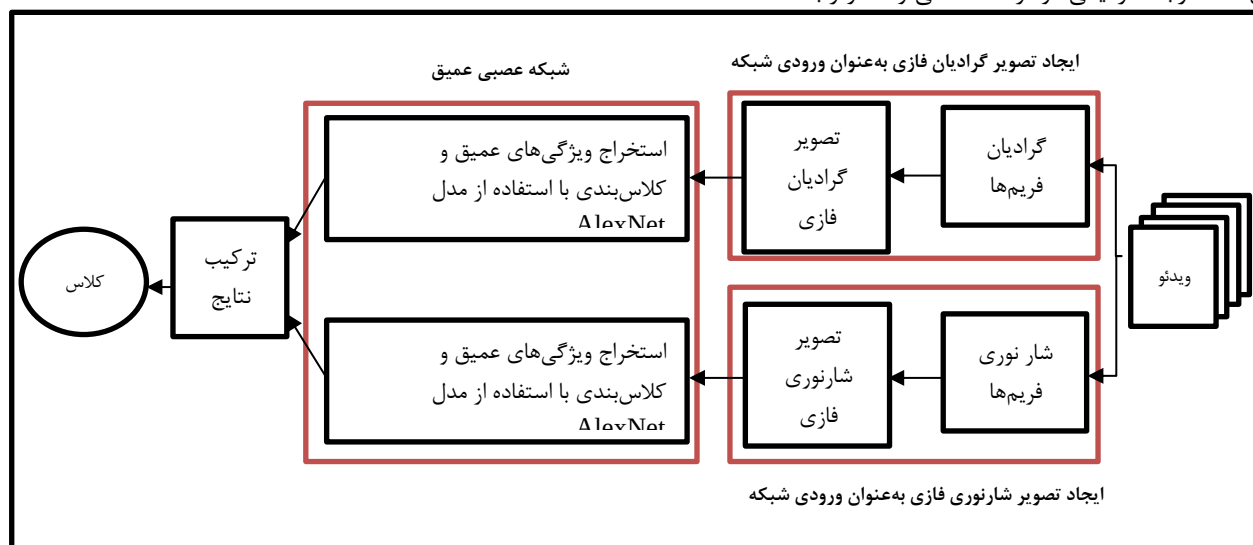
و برای فریم‌های رنگی گرادیان هر کانال جداگانه محاسبه می‌شود.

۳-۳- شار نوری

مفهوم شار نوری در سال ۱۹۴۰ توسط گیبسون [۳۵] برای توصیف محرک‌های دیداری فراهم شده است. اهمیت شار نوری به دلیل قابلیت درک و توانایی تشخیص حالات ممکن برای انجام یک عمل مورد توجه قرار گرفته است. شار نوری یک روش مناسب برای استخراج اطلاعات حرکتی از دنباله فریم‌ها است. استفاده از شار نوری روی فریم‌های پی‌درپی می‌تواند حرکت هر پیکسل را در جهت عمودی و افقی استخراج کند. یک توصیف مناسب از همه این حرکات در هر پیکسل می‌تواند در کاربردهای گوناگون متفاوت باشد [۳۶].

روش‌های شار نوری سعی بر محاسبه حرکت بین دو فریم در زمان‌های t و $t + \Delta t$ را دارند. برای نمونه برای یک پیکسل در موقعیت (x, y, t) با شدت پیکسل $I(x, y, t)$ به اندازه‌های v_x و v_y بین دو فریم پی‌درپی از تصاویر حرکت خواهد کرد و محدودیت پایداری روشی به صورت زیر است:

دقیق تخمین زد، به همین دلیل استفاده از رابطه فازی مکانی فاصله که نشان‌دهنده رابطه نزدیکی دو فرد است، می‌تواند مؤثر باشد.



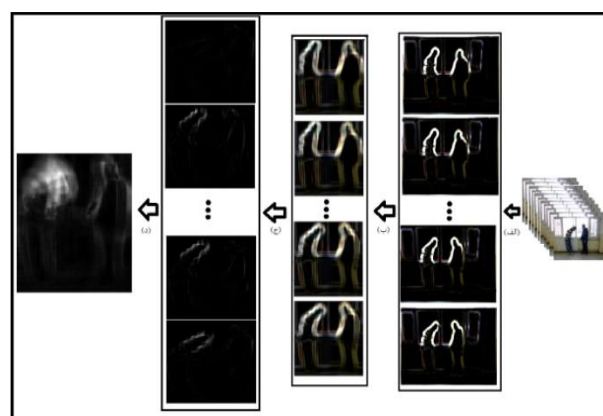
شکل ۳: فلوچارت روش ارائه‌شده

$$\mu_1(i, j) = \begin{cases} 0 & j \leq 0 \\ \frac{j-a}{b-a} & a \leq j \leq b \\ 1 & b < j \leq c \\ \frac{d-j}{d-c} & c \leq j \leq d \\ 0 & d \leq j \end{cases} \quad (5)$$

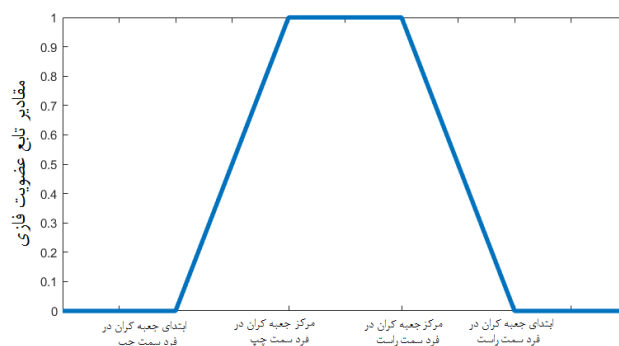
که مقدار ستون ابتدای جعبه کران فرد سمت چپ، b مقدار ستون مرکز جعبه کران فرد سمت چپ، c مقدار ستون ابتدای جعبه کران فرد سمت راست، d مقدار ستون مرکز جعبه کران فرد سمت راست است. در شکل ۴ (ب)، تابع عضویت فاصله بر تصاویر اعمال می‌شود تا ناحیه بین دو فرد استخراج شود. در مرحله بعد اختلاف اندازه تصاویر گرادیان این ناحیه هر فریم با فریم بعدی محاسبه‌شده است (شکل ۴(ج)).

از طرفی در هنگام انجام تعامل در ویدئوهای مختلف زمان انجام آن یکسان نخواهد بود، به همین دلیل برای اهمیت‌دادن به فریم‌هایی که احتمال بیشتری دارند که تعامل در حال اجرا باشد از یک تابع عضویت فازی بر مبنای فاصله زمانی فریم‌ها استفاده می‌شود تا وزن درستی به هر فریم نسبت داده شود (شکل ۶). هنگام انجام تعامل در فریم‌های وسط ویدئو احتمال انجام تعامل بیشتر است. به همین دلیل یک تابع عضویت که دارای سه حالت (آغاز، وسط، انتها) تعریف شده، که بر اساس شروع ویدئو تا انتهای آن است. در فریم‌های میانی ارزش پیکسل‌ها برای ایجاد تصویر فازی بیشتر است، از این تابع برای جمع فریم‌ها استفاده خواهد شد. از این تابع عضویت در مقاله [۲۳] نیز استفاده شده است.

$$\mu_2(i) = \begin{cases} \frac{2i}{n} & 0 < i \leq n/2 \\ 2 - \frac{2i}{n} & n/2 < i \leq n \end{cases} \quad (6)$$



شکل ۴: ایجاد تصویر فازی بر مبنای گرادیان؛ (الف) محاسبه گرادیان هر فریم، (ب) اعمال تابع عضویت فاصله μ_1 بر تصویر گرادیان، (ج) محاسبه اختلاف تصاویر گرادیان هر فریم با فریم بعد آن، (د) ایجاد تصویر فازی با استفاده از تابع عضویت μ_2 .

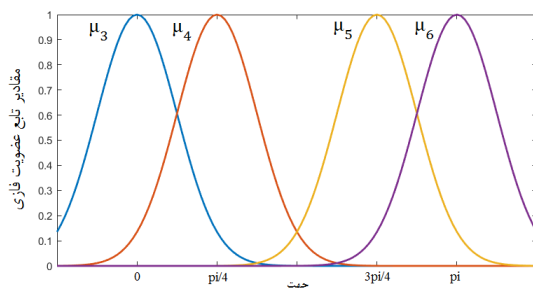


شکل ۵: تابع عضویت فاصله

در شکل ۵ تابع عضویت برای فاصله نشان داده شده است، که برای هر پیکسل (i, j) در فریم مقدار $\mu_1(i, j)$ از رابطه (۵) به دست می‌آید.

$$\mu_{3,4,5,6}(i,j) = e^{\frac{-(\theta_{ij}-c)^2}{2\sigma^2}} \quad (7)$$

چهار تابع عضویت بر اساس رابطه (۷) تشکیل شده‌اند که در شکل ۸ نشان داده شده است. مرکز توابع به ترتیب $0, \pi/4, 3\pi/4, \pi$ است. مقدار σ برابر $\pi/8$ است. مقدار θ_{ij} برابر جهت پیکسل (i,j) در تصویر شار نوری است.

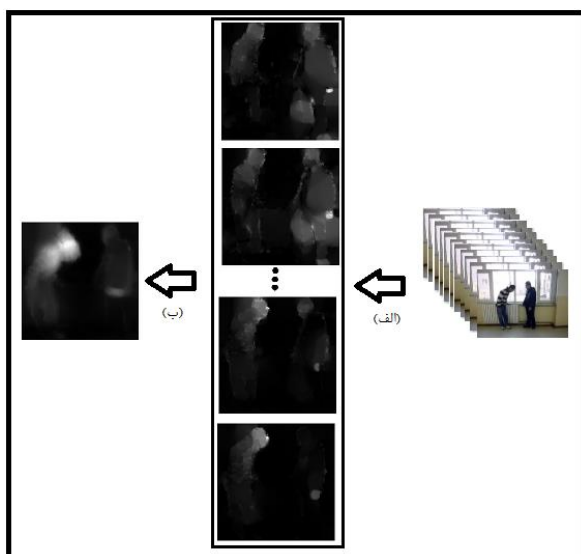


شکل ۸: تابع عضویت مکانی

برای هر پیکسل مقدار اندازه شارنوری فازی برابر مقدار زیر است:

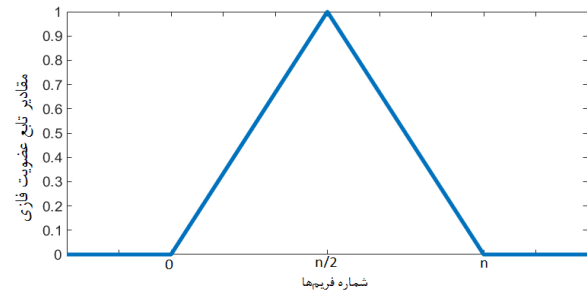
$$\rho(i,j) = \mu \sqrt{v(i,j)_x^2 + v(i,j)_y^2} \quad (8)$$

که مقدار μ برای هر پیکسل از تابع عضویت شکل ۸ به دست می‌آید. برای حالتی که پیکسلی مقدار $\mu_3, \mu_4, \mu_5, \mu_6$ در بیش از یک حالت غیر صفر است، ماکزیمم آن در نظر گرفته می‌شود. از مجموع وزن دار تصاویر شار نوری تصویری ایجاد می‌شود که برای تشخیص تعامل استفاده می‌شود. چون در فریم‌های وسط ویدئو احتمال انجام تعامل بیشتر است، فریم‌های وسط ارزش بیشتری دارد و از تابع عضویت شکل ۶ برای وزن دار کردن فریم‌ها استفاده خواهد شد.



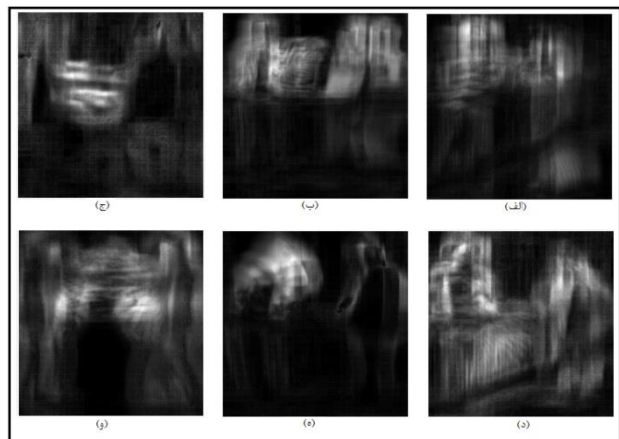
در معادله (۶)، n تعداد فریم‌ها در ویدئو است. $\mu_2(i)$ نشان دهنده مقدار تابع عضویت در i امین فریم ویدئو است.

در شکل ۴ (د)، ترکیب وزن دار فریم‌ها بعد از اعمال تابع عضویت μ_2 نشان داده شده است. در این مرحله تصویر گرادیان فازی برای ورودی به شبکه ایجاد خواهد شد.



شکل ۶: تابع عضویت فاصله زمانی فریم‌ها، n نشان دهنده تعداد فریم‌ها در ویدئو است.

در شکل ۷ نمونه‌هایی از تصویر فازی بر مبنای گرادیان نشان داده شده است.



شکل ۷: نمونه‌ای از تصاویر گرادیان فازی به دست آمده: (الف) هل دادن، (ب) مشت زدن، (ج) دست دادن، (د) لگد زدن، (ه) تعظیم کردن، (و) بزن قدش.

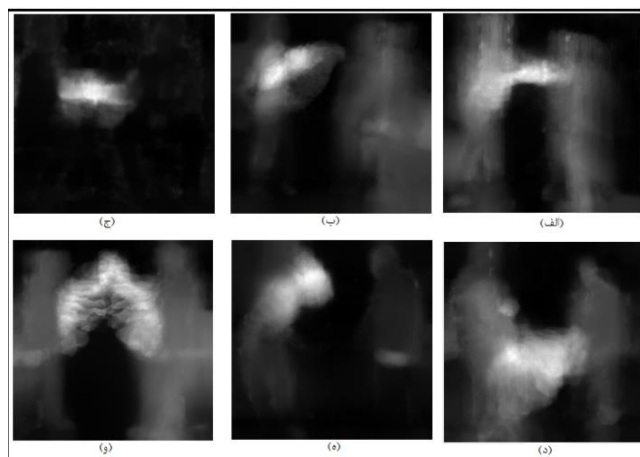
۴-۲- ایجاد تصویر شار نوری فازی

برای ایجاد تصویر شار نوری فازی، بعد از پیدا کردن جعبه کران هر دو فرد، جهت حرکت آن‌ها بررسی می‌شود. چون فرد در حال تعامل، یک نفر از چپ به راست و فرد دیگر از راست به چپ حرکت می‌کند، یک تابع عضویت بر اساس رابطه مکانی آن‌ها تعریف می‌شود. در فرد سمت چپ حرکت به سمت راست و فرد سمت راست بالعکس اهمیت دارد و حرکت بقیه پیکسل‌ها برای تشخیص تعامل اهمیتی ندارد، پس نقاطی که جهت آن به سمت مورد نظر است، بررسی می‌شود.

تابع عضویت در شکل ۸ برای ارزش گذاری جهت حرکت‌های مؤثر در تشکیل نوع تعامل ارائه شده است. برای فرد سمت راست دو تابع μ_4, μ_3 و برای فرد سمت چپ دو تابع μ_5, μ_6 استفاده شده است. از جهت تصاویر شار نوری برای محاسبه جهت حرکت استفاده شده است.

شکل ۹: ایجاد تصویر فازی بر مبنای شارنوری؛ (الف) محاسبه شارنوری هر فریم با استفاده از توابع عضویت $\mu_{3,4,5,6}$ ، (ب) ایجاد تصویر فازی با استفاده از تابع عضویت μ_2 .

در شکل ۹ مراحل ایجاد تصویر فازی شارنوری نشان داده شده است. همان‌طور که در قسمت (الف) این شکل ۹ نشان داده شده ابتدا تصویر شارنوری فازی برای هر فریم ایجاد، سپس در قسمت (ب) یک تصویر کامل از تمام تصاویر مرحله قبل ایجاد شده است. در شکل ۱۰ نمونه‌هایی از تصویر فازی بر مبنای شارنوری نشان داده شده است. این تصاویر نشان می‌دهد تصویر فازی ایجاد شده دارای اطلاعات حرکتی متمایز برای تشخیص نوع تعامل است.



شکل ۱۰: نمونه‌ای از تصاویر شارنوری فازی به دست آمده؛ (الف) هل دادن، (ب) مشت زدن، (ج) دست دادن، (د) لگزدن، (ه) تعظیم کردن، (و) بزن قدش.

۴-۳- شبکه عصبی عمیق

در بخش قبل فرآیند محاسبه تصویر فازی توضیح داده شد، مرحله بعد استخراج ویژگی از این تصاویر است. انتخاب ویژگی مهم‌ترین بخش برای نمایش ویژگی‌های مخفی و بصری داده است. به همین دلیل برای استخراج ویژگی از تصاویر فازی از یک شبکه کانولوشن استفاده شده است. در این کار از معماری AlexNet [۵] که بر روی ILSVRC-2012 [۳۹] آموزش داده شده، استفاده شده است. دو تصویر فازی استخراج شده از مرحله قبل هر کدام به یک شبکه CNN، از پیش آموزش داده شده، ارسال شده است. در مرحله آخر شبکه یک لایه softmax قرار داده شده که احتمال هر تصویر به کلاس‌های تعامل محاسبه می‌شود. دو خروجی از دو شبکه در یکدیگر ضرب شده و کلاسی که احتمال بیشتری دارد به عنوان کلاس تصاویر ورودی، نسبت داده می‌شود. ترکیب به صورت زیر انجام می‌شود [۴۰]:

$$label = F_{max}(v_1 \circ v_2) \quad (9)$$

V بردار امتیازات است، $\circ$ عملگر ضرب است و F_{max} تابعی است که ایندکسی که بیشترین مقدار را دارد، برمی‌گرداند.

۵- آزمایشات

در این بخش روش پیشنهادی ارزیابی و عملکرد آن با روش‌های دیگر مقایسه شده است. در بخش ۵-۱ دو مجموعه داده برای ارزیابی معرفی شده است. جزییات پیاده سازی در بخش ۵-۲- شرح داده شده است. در بخش ۵-۳ نتایج روش و تحلیل آن بیان شده است.

۵-۱- مجموعه داده

روش ارائه شده بر روی دو مجموعه داده تعامل که به طور گسترده بررسی و استفاده می‌شود به نام مجموعه داده تعامل BIT و UT چک شده است.

مجموعه داده تعامل BIT: این مجموعه داده شامل ۸ نوع تعامل (تعظیم کردن، مشت زدن، دست دادن، بزن قدش، بغل کردن، لگزدن، نوازش کردن، هل دادن) بین دو انسان است [۴۱]. هر کلاس شامل ۵۰ ویدیو است. این مجموعه داده به خاطر تغییرات نور، تفاوت ظاهر و سبزه افراد، نقطه دید و علاوه بر آن تصادم به دلیل وجود افراد دیگر، پل‌ها و .. بسیار چالش برانگیز است.

مجموعه داده تعامل UT: این مجموعه شامل دو دسته است [۴۲]. پیش زمینه دسته اول ساده و ایستاست. برخلاف آن مجموعه دوم دارای پیش زمینه پیچیده تر با حرکت ملایم است. هر مجموعه شامل ۶۰ ویدیو و شامل ۶ کنش دست دادن، بغل کردن، اشاره کردن، لگزدن، هل دادن و مشت زدن است. هر ویدیوی تست تعامل را در ده نسبت مشاهده پیش بینی می‌کند. به عبارت دیگر ویدیو به ۱۰ قسمت تقسیم می‌شود.

۵-۲- جزییات پیاده سازی

برای انتخاب توابع عضویت استفاده در این مقاله، توابع گوسی، مثلثی و دوزنقه مورد بررسی قرار گرفت و با استفاده از آزمون هر تابع و محدوده‌های مختلف بهترین تابع با نظر افراد خبره برای هر حالت انتخاب شد. شبکه ارائه شده روی چهارچوب کافه [۴۳] پیاده سازی شده است. نرخ یادگیری برای تنظیم وزن‌ها ۰/۰۰۱ و پارامتر تنزل وزن ۰/۰۱ است. مقدار تعداد تکرار برای تنظیم وزن‌ها برابر ۱۰۰۰۰ تکرار است. در فاز تست، هر ویدیو به ۱۰ نرخ مشاهده از ۱/ تا یک تقسیم می‌شود. به طور مثال اگر ویدیویی n فریم داشته باشد و نرخ مشاهده p باشد $p \times n$ فریم اول برای استخراج تصویر فازی و پیش بینی تعامل استفاده خواهد شد.

۵-۳- نتایج و بحث

ابتدا نتایج دقت شبکه با ورودی تصویر گرادین فازی و تصویر شارنوری فازی به صورت مجزا و ترکیب خروجی دو شبکه بر روی مجموعه داده BIT مورد بررسی قرار می‌گیرد. همان‌طور که در جدول ۱ نشان

تابع ϵ ضویت فاصله زمانی در نظر گرفته نمی شود، دقت شبکه پایین می آید.

جدول ۲: دقت روش ارائه شده قبل از اعمال تابع عضویت μ_1 و μ_2

دقت	ورودی
۸۵/۱۶	قبل از اعمال تابع عضویت μ_1
۸۲/۰۳	بعد از اعمال تابع عضویت μ_2
۸۹/۸۴	بعد از اعمال تابع عضویت μ_1 و μ_2

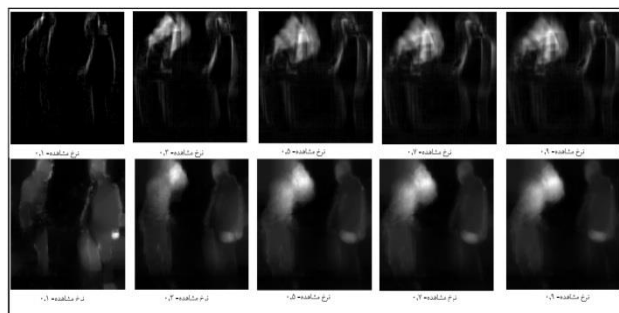
در تصویر شارنوری اختلاف پیکسل ها در زمان محاسبه می شود، اما در تصویر گرادیان، اختلاف بین گرادیان ها محاسبه می شود.

در شکل ۱۲ ماتریس درهم ریختگی برای مجموعه داده BIT نشان داده شده است. همان طور که نشان داده شده در حالتی که کنش ها شبیه هم هستند به خاطر شباهت الگو حرکتی آن ها روش ارائه شده قادر به تشخیص نبوده است.

	دست دادن	بزن قدش	یغل کردن	تعظیم کردن	مشت زدن	لگد زدن	نوازش کردن	هل دادن
دست دادن	-	-	-	-	-	-	-	-
بزن قدش	-	-	-	-	-	-	-	-
یغل کردن	-	-	-	-	-	-	-	-
تعظیم کردن	-	-	-	-	-	-	-	-
مشت زدن	-	-	-	-	-	-	-	-
لگد زدن	-	-	-	-	-	-	-	-
نوازش کردن	-	-	-	-	-	-	-	-
هل دادن	-	-	-	-	-	-	-	-

شکل ۱۲: ماتریس درهم ریختگی بر روی مجموعه BIT

نمونه ای از تصاویر گرادیان فازی و شارنوری فازی مربوط به تعامل تعظیم کردن در نرخ های مشاهده مختلف در شکل ۱۳ نشان داده شده است.



شکل ۱۳: نمونه ای از تصاویر مربوط به تعامل تعظیم کردن از مجموعه داده BIT. سطر اول: تصاویر گرادیان فازی با نرخ ۰.۱، ۰.۳، ۰.۵، ۰.۷، ۰.۹. سطر دوم: تصاویر شارنوری فازی با نرخ ۰.۱، ۰.۳، ۰.۵، ۰.۷، ۰.۹.

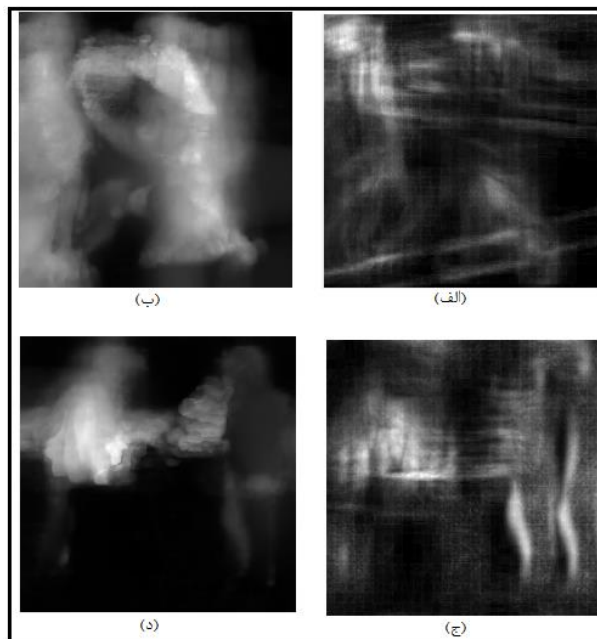
این روش را با سایر روش های تشخیص و پیش بینی تعامل مقایسه شده است. نتایج در شکل ۱۴ (الف) و جدول ۳، نشان می دهد روش ارائه شده در ۵ نرخ مشاهده از ۱۰ نرخ بهترین دقت را داشته است.

داده شده ترکیب نتایج با استفاده معادله (۶) نتایج بهتری را داشته است. در جدول ۱ نتایج بر اساس ترکیب نتایج این دو ورودی ارائه خواهد شد.

جدول ۱: دقت روش ارائه شده قبل از ترکیب نتایج شبکه

دقت	ورودی
۸۵/۹۴	تصویر گرادیان فازی
۸۰/۴۷	تصویر شارنوری فازی
۸۹/۸۴	هر دو تصویر

همان طور که در شکل ۱۱ نشان داده شده، تصاویر دو تعامل نشان داده شده است. تصویر (الف) و (ب) به ترتیب تصویر گرادیان فازی و تصویر شارنوری فازی مربوط به مشت زدن است که خروجی شبکه تصویر (ب) را به درستی مشت زدن و تصویر (الف) را به اشتباه هل دادن برچسب زده است، اما در هنگام ترکیب این دو نتیجه مشت زدن برگردانده شده است. در سطر دوم نیز تصویر (ج) و (د) به ترتیب تصویر گرادیان فازی و تصویر شارنوری فازی مربوط به تعامل بزن قدش است، خروجی شبکه برای تصویر (ج) به درستی تشخیص داده شده اما تصویر (د) را تعامل دست دادن برچسب زده است. در این حالت نیز ترکیب نتیجه دو شبکه نتیجه درست برگردانده است.



شکل ۱۱: سطر اول مربوط به تعامل مشت زدن؛ (الف) تصویر گرادیان، (ب) تصویر فازی شارنوری آن؛ سطر دوم مربوط به تعامل بزن قدش، (ج) تصویر گرادیان، (د) تصویر شارنوری.

در جدول ۲ نتایج مربوط به عدم اعمال تابع عضویت μ_1 و μ_2 نشان داده شده است. این نتایج بر روی مجموعه داده BIT و بر روی ساختار ارائه شده روش پیشنهادی است. در حالت اول تمام توابع عضویت به جز μ_1 و در حالت دوم تمام توابع عضویت به جز μ_2 اعمال شده است. همان طور که در جدول ۲ نشان داده شده، هنگامی که

به طور مثال هنگامی که ۰/۷ فریم‌های یک ویدئو دیده شده حدود ۵ در صد روش ارائه شده دقت بهتری داشته است. در حالتی که نرخ بین ۰/۲ تا ۰/۵ از ویدئو مشاهده شده، روش ارائه دقت خیلی پایینی نسبت به روش‌های قبلی نداشته و تقریباً به بقیه روش‌ها نزدیک است. نتایج این روش بر روی مجموعه داده UT نیز بررسی شده است. همان‌طور که در بخش ۵-۱ ذکر شد، این مجموعه شامل دو دسته است. نتایج بر روی دسته اول در

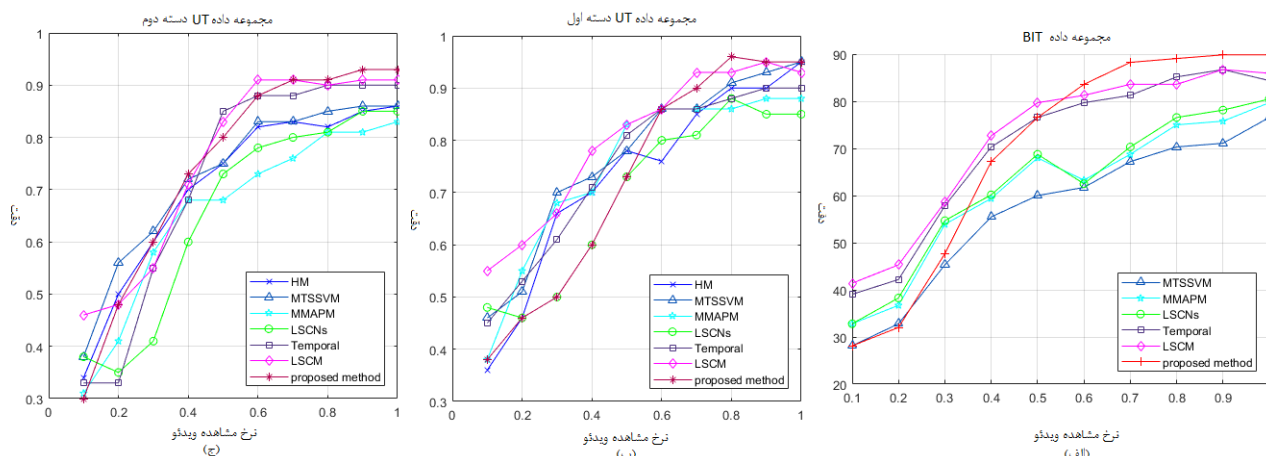
دلیل این مسئله این است که تصویر فازی ایجاد شده به دلیل جابه‌جایی کم و آغازی یک کنش اطلاعات حرکتی زیادی ندارد اما هر چه به سمت انجام کنش حرکت می‌کند تصاویر اطلاعات حرکتی بیشتری خواهند داشت.

روش پیشنهادی و سایر روش‌ها هنوز توانسته‌اند در نرخ‌های مشاهده بالا (۰/۹ و یک) به دقت ۱۰۰٪ برسند، این به دلیل چالش‌های موجود مانند سرعت حرکت افراد و نحوه اجرای متفاوت آنان در انجام یک تعامل یکسان است.

روش LSCM نسبت به روش‌های ارائه شده قبلی نتیجه بهتری دارد این روش بر مبنای شبکه LSTM است و نیاز به محاسبات بالایی دارد اما روش ما توانسته با محاسبات کمتر نتایج مورد قبولی داشته باشد. در این حالت فقط یک تصویر به شبکه ارسال می‌شود درحالی‌که در بقیه روش‌ها تصویر فریم به فریم ویدئو باید به شبکه ارسال شود.

جدول ۴، شکل ۱۴ (ب) و نتایج بر دو دسته دوم در جدول ۵، شکل ۱۴ (ج) نشان داده شده است. روش ارائه شده بر روی دسته اول در نرخ مشاهده ۰/۶، ۰/۸، ۰/۹ و یک، و بر روی دسته دوم در نرخ مشاهده ۰/۴، ۰/۷، ۰/۸، ۰/۹ و یک، نتایج بهتری داشته است. نتایج در سایر نرخ‌های مشاهده تقریباً مانند روش‌های دیگر بوده و دقت خوبی داشته است.

همان‌طور که نتایج نشان می‌دهد، دقت روش در نرخ‌های پایین افزایش چندانی نسبت به روش‌های دیگر نداشته است.



شکل ۱۴: مقایسه دقت مجموعه داده: (الف) BIT، (ب) UT دسته اول، (ج) UT دسته دوم، با سایر روش‌ها.

جدول ۳: نتایج دقت پیش‌بینی تعامل روی مجموعه داده BIT با نرخ‌های مشاهده متفاوت

نرخ مشاهده										روش
۱	۰/۹	۰/۸	۰/۷	۰/۶	۰/۵	۰/۴	۰/۳	۰/۲	۰/۱	
۷۶/۵۶	۷۱/۰۹	۷۰/۳۱	۶۷/۱۹	۶۱/۷۲	۶۰	۵۵/۴۷	۴۵/۳۱	۳۲/۸۱	۲۸/۱۳	MTSSVM[13]
۷۹/۶۹	۷۵/۷۸	۷۵	۶۸/۷۵	۶۳/۲۸	۶۷/۹۷	۵۹/۳۸	۵۳/۹۰	۳۶/۷۲	۳۲/۸۱	MMAPM[14]
۸۰	۷۸	۷۶	۷۰	۶۲	۶۸	۶۰	۵۴	۳۸	۳۲	LSCNs[28]
۸۴/۳۸	۸۶/۷۲	۸۵/۱۶	۸۱/۲۵	۷۹/۶۹	۷۶/۵۶	۷۰/۳۰	۵۷/۸۱	۴۲/۱۹	۳۹/۰۶	Temporal[26]
۸۵/۹۴	۸۶/۷۲	۸۳/۵۹	۸۳/۵۹	۸۱/۲۵	۷۹/۶۹	۷۲/۶۶	۵۸/۵۹	۴۵/۳۱	۴۱/۴۱	LSCM[27]
۸۹/۸۴	۸۹/۸۴	۸۹/۰۶	۸۸/۲۵	۸۳/۵۹	۷۶/۵۶	۶۷/۱۹	۴۷/۶۶	۳۲/۰۳	۲۸/۱۳	روش ارائه شده

جدول ۴: نتایج دقت پیش‌بینی تعامل روی مجموعه داده دسته اول UT با نرخ‌های مشاهده متفاوت

نرخ مشاهده	روش									
	۰/۱	۰/۲	۰/۳	۰/۴	۰/۵	۰/۶	۰/۷	۰/۸	۰/۹	۱
HM[11]	۰/۳۸	۰/۵۵	۰/۶۸	۰/۷	۰/۸۳	۰/۸۶	۰/۸۶	۰/۸۶	۰/۸۸	۰/۸۸
MTSSVM[13]	۰/۳۶	۰/۴۶	۰/۶۶	۰/۷	۰/۷۸	۰/۷۶	۰/۸۵	۰/۹	۰/۹	۰/۹۵
MMAPM[14]	۰/۴۶	۰/۵۱	۰/۷۰	۰/۷۳	۰/۷۸	۰/۸۶	۰/۸۶	۰/۹۱	۰/۹۳	۰/۹۵
LSCNs[28]	۰/۴۸	۰/۴۶	۰/۵	۰/۶	۰/۷۳	۰/۸	۰/۸۱	۰/۸۸	۰/۸۵	۰/۸۵
Temporal[26]	۰/۴۵	۰/۵۳	۰/۶۱	۰/۷۱	۰/۸۱	۰/۸۶	۰/۸۶	۰/۸۸	۰/۹	۰/۹
LSCM[27]	۰/۵۵	۰/۶	۰/۶۶	۰/۷۸	۰/۸۳	۰/۸۶	۰/۹۳	۰/۹۳	۰/۹۵	۰/۹۳
روش ارائه شده	۰/۳۸	۰/۴۶	۰/۵	۰/۶	۰/۷۳	۰/۸۶	۰/۹	۰/۹۶	۰/۹۵	۰/۹۵

جدول ۵: نتایج دقت پیش‌بینی تعامل روی مجموعه داده دسته دوم UT با نرخ‌های مشاهده متفاوت

نرخ مشاهده	روش									
	۰/۱	۰/۲	۰/۳	۰/۴	۰/۵	۰/۶	۰/۷	۰/۸	۰/۹	۱
HM[11]	۰/۳۱	۰/۴۱	۰/۵۸	۰/۶۸	۰/۶۸	۰/۷۳	۰/۷۶	۰/۸۱	۰/۸۱	۰/۸۳
MTSSVM[13]	۰/۳۴	۰/۵	۰/۶	۰/۷	۰/۷۵	۰/۸۲	۰/۸۳	۰/۸۲	۰/۸۵	۰/۸۶
MMAPM[14]	۰/۳۸	۰/۵۶	۰/۶۲	۰/۷۲	۰/۷۵	۰/۸۳	۰/۸۳	۰/۸۵	۰/۸۶	۰/۸۶
LSCNs[28]	۰/۳۸	۰/۳۵	۰/۴۱	۰/۶	۰/۷۳	۰/۷۸	۰/۸	۰/۸۱	۰/۸۵	۰/۸۵
Temporal[26]	۰/۳۳	۰/۳۳	۰/۵۵	۰/۶۸	۰/۸۵	۰/۸۸	۰/۸۸	۰/۹	۰/۹	۰/۹
LSCM[27]	۰/۴۶	۰/۴۸	۰/۵۵	۰/۷۱	۰/۸۳	۰/۹۱	۰/۹۱	۰/۹۱	۰/۹۱	۰/۹۱
روش ارائه شده	۰/۳	۰/۴۸	۰/۶	۰/۷۳	۰/۸	۰/۸۸	۰/۹۱	۰/۹۱	۰/۹۳	۰/۹۳

مراجع

۶- نتیجه‌گیری

در این مقاله یک شبکه عصبی عمیق برای پیش‌بینی تعامل با استفاده از روابط فازی و شار نوری ارائه شده است. در این روش ابتدا یک تصویر فازی بر مبنای گرادینان و یک تصویر فازی بر مبنای شار نوری ایجاد شده، سپس با استفاده از شبکه کانولوشن از این تصاویر ویژگی استخراج شده و در نهایت خروجی شبکه با یکدیگر ترکیب شده است. تصویر فازی ارائه شده دارای اطلاعات حرکتی مناسب برای تشخیص و پیش‌بینی تعامل است.

این تصویر ترکیبی از اطلاعات حرکتی مناسب در زمان‌های انجام فرآیند تعامل است، به همین دلیل به جای بررسی ویژگی‌های تک‌تک فریم‌های یک ویدئو فقط اطلاعات مناسب استخراج می‌شود. در روش پیشنهادی بر محدودیت بالا بودن کنتراست فریم‌های رنگی برای استخراج ویژگی‌های ظاهری از تصاویر غلبه شده است.

در تصویر شارنوری اختلاف پیکسل‌ها در زمان، در تصویر گرادینان، اختلاف بین گرادینان‌ها محاسبه می‌شود، ترکیب نتایج دسته‌بندی این دو تصویر باعث بهبود دقت شده است. نتایج بر روی دو مجموعه داده استاندارد UT و BIT نشان داده که روش ارائه شده دقت قابل قبولی داشته است. در آینده هدف توسعه این روش را برای پیش‌بینی تعامل در فعالیت‌های گروهی و پیش‌بینی کنش‌های فردی است.

- [1] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," IEEE International Conference on Computer Vision, 2011.
- [2] M. Ramanathan, W. Y. Yau and E. K. Teoh, "Human action recognition with video data: research and evaluation challenges," IEEE Transactions on Human-Machine Systems, vol. 44, no. 5, pp. 650-663, 2014.
- [3] N. G. Cho, S. H. Park, J. S. Park, U. Park and S. W. Lee, "Compositional interaction descriptor for human interaction recognition," Neurocomputing, vol. 265, pp. 169-181, 2017.

[۴] امیر سزاوار، حسن فرسی، سجاد محمدزاده، «بازیابی تصویر مبتنی بر

محتوا با استفاده از شبکه‌های عصبی کانولوشن عمیق»، مجله مهندسی

برق دانشگاه تبریز، جلد ۴۸، شماره ۴، صفحه ۱۶۰۳-۱۵۹۵، زمستان

۱۳۹۷.

- [5] A. Krizhevsky, I. Sutskever and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, pp.1097-1105, 2012.
- [6] G. Acampora, P. Foggia, A. Saggese and M. Vento, "Combining neural networks and fuzzy systems for human behavior understanding," IEEE International Conference on Advanced Video and Signal-based Surveillance, 2012.
- [7] J. Y. Chang, J. J. Shyu, and C. W. Cho, "Fuzzy rule inference based human activity recognition," IEEE Conference on Control Applications and IEEE International Symposium on Intelligent Control, 2009.
- [8] H. Medjahed, D. Istrate, J. Boudy and B. Dorizzi, "Human activities of daily living recognition using fuzzy logic for elderly home monitoring," IEEE International Conference on Fuzzy Systems, 2009.

- weighting,” *Contemp. Eng. Sci.*, vol. 10, no. 29, pp. 1419-1429, 2017.
- [26] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, “Human interaction prediction using deep temporal features,” *European Conference on Computer Vision*, Springer, 2016.
- [27] Q. Ke, M. Bennamoun, S. An, F. Sohel and F. Boussaid, “Leveraging structural context models and ranking score fusion for human interaction prediction,” *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1712-1723, 2018.
- [28] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [29] A. Stergiou and R. Poppe, “Understanding human-human interactions: a survey,” *arXiv preprint arXiv:1808.00022*, 2018.
- [30] X. Wang and J.M. Keller, “Human-based spatial relationship generalization through neural/fuzzy approaches,” *Fuzzy Sets and Systems*, vol. 101, no. 1, pp. 5-20, 1999.
- [31] R. Pierrard, J. P. Poli, and C. Hudelot, “Learning Fuzzy Relations and Properties for Explainable Artificial Intelligence,” *IEEE International Conference on Fuzzy Systems*, 2018.
- [32] H. Hüttenrauch, K. S. Eklundh, A. Green and E. A. Topp, “Investigating spatial relationships in human-robot interaction,” *International Conference on Intelligent Robots and Systems*, 2006.
- [33] I. Bloch, “Fuzzy spatial relationships for image processing and interpretation: a review,” *Image and Vision Computing*, vol. 23, no. 2, pp. 89-110, 2005.
- [34] A. Delmonte, I. Bloch, D. Hasboun, C. Mercier, J. Pallud and P. Gori, “Segmentation of white matter tractograms using fuzzy spatial relations,” *Organization for Human Brain Mapping*, 2018.
- [35] J. J. Gibson, *The perception of the visual world*, 1950.
- [36] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” *European conference on Computer Vision*, Springer, 2004.
- [37] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black and R. Szeliski, “A database and evaluation methodology for optical flow,” *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1-31, 2011.
- [38] A. Ess, B. Leibe, K. Schindler and L. Van Gool, “A mobile vision system for robust multi-person tracking,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma and A. C. Berg, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.
- [40] C. Li, Y. Hou, P. Wang, and W. Li, “Joint distance maps based action recognition with convolutional neural networks,” *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624-628, 2017.
- [41] Y. Kong, Y. Jia, and Y. Fu., “Learning human interaction by interactive phrases,” *European Conference on Computer Vision*, Springer, 2012.
- [42] M. S. Ryoo and J. Aggarwal, “UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA),” *IEEE International Conference on Pattern Recognition Workshops*, 2010.
- [43] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *ACM International Conference on Multimedia*, 2014.
- [۹] ندا خانبانی، امیرمسعود افتخاری مقدم، «ارائه یک روش تشخیص زبان علامت مبتنی بر رویکرد MLRF فازی با استفاده از اطلاعات عمق تصویر»، *مجله مهندسی برق دانشگاه تبریز*، ۱۳۹۶، *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۷، شماره ۳، صفحه ۹۸۷-۹۷۷، پاییز ۱۳۹۶.
- [10] N. P. Trong, H. Nguyen, K. Kazunori and B. Le Hoai, “A comprehensive survey on human activity prediction,” *International Conference on Computational Science and Its Applications*, Springer, 2017.
- [11] T. Lan, T. C. Chen, and S. Savarese, “A hierarchical representation for future action prediction,” *European Conference on Computer Vision*, Springer, 2014.
- [12] C. Gao, L. Yang, Y. Du, Z. Feng and J. Liu, “From constrained to unconstrained datasets: an evaluation of local action descriptors and fusion strategies for interaction recognition,” *World Wide Web*, vol. 19, no. 2, pp. 265-276, 2016.
- [13] Y. Kong, D. Kit, and Y. Fu, “A discriminative model with multiple temporal scales for action prediction,” *European Conference on Computer Vision*, Springer, 2014.
- [14] Y. Kong and Y. Fu, “Max-margin action prediction machine,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1844-1858, 2016.
- [15] W. Choi, K. Shahid, and S. Savarese, “What are they doing?: Collective activity classification using spatio-temporal relationship among people,” *IEEE International Conference on Computer Vision, Workshops*, pp. 1282-1289, 2009.
- [16] Z. Wang, S. Liu, J. Zhang, S. Chen, and Q. Guan, “A spatio-temporal CRF for human interaction understanding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1647-1660, 2017.
- [17] J. M. Le Yaouanc and J.-P. Poli, “A fuzzy spatio-temporal-based approach for activity recognition,” *International Conference on Conceptual Modeling*, Springer, 2012.
- [18] B. Yao, H. Hagnas, M. J. Alhaddad and D. Alghazzawi, “A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments,” *Soft Computing*, vol. 19, no. 2, pp. 499-506, 2015.
- [19] K. Mozafari, N. M. Charkari, H. S. Boroujeni and M. Behrouzifar, “A novel fuzzy hmm approach for human action recognition in video,” *Knowledge Technology*, Springer, pp. 184-193, 2012.
- [20] M. Raptis and L. Sigal, “Poselet key-framing: A model for human activity recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [21] A. Iosifidis, A. Tefas, and I. Pitas, “Activity-based person identification using fuzzy representation and discriminant learning,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 530-542, 2012.
- [22] L. Nanni, S. Ghidoni, and S. Brahnam, “Handcrafted vs. non-handcrafted features for computer vision classification,” *Pattern Recognition*, vol. 71, pp. 158-172, 2017.
- [23] E. P. Ijjina and K. M. Chalavadi, “Human action recognition in RGB-D videos using motion sequence information and deep learning,” *Pattern Recognition*, vol. 72, pp. 504-516, 2017.
- [24] H. J. Kim, J. S. Lee, and H. S. Yang, “Human action recognition using a modified convolutional neural network,” *International Symposium on Neural Networks*, Springer, 2007.
- [25] R. J. Moreno, O. A. Sanchez, and D. M. Ovalle, “RGB-D training for convolutional neural network with final fuzzy layer for depth

زیر نویس ها

- ¹⁰ Hierarchical Movemes
- ¹¹ Support Vector Machine
- ¹² Scale-invariant feature transform
- ¹³ Multiple Temporal Scale Support Vector Machine
- ¹⁴ Bag of Word
- ¹⁵ Max-Margin Action Prediction Machine
- ¹⁶ silhouette
- ¹⁷ handcrafted
- ¹⁸ Caffe
- ¹⁹ Accuracy

- ¹ Motion Energy Image
- ² Motion History Image
- ³ Convolutional Neural Network
- ⁴ Optical Flow
- ⁵ Long short-term memory
- ⁶ Histogram of Gradient
- ⁷ Histogram of Optical Flow
- ⁸ Motion Boundary Histogram
- ⁹ Space Time Interest Points