:STHARNet تشخیص شبکه عمل مکانی-زمانی انسان شبکه در بازیابی ویدئو مبتنی بر محتوا

**چکیده:**

بیشتر اطلاعات مورد نیاز با استفاده از اینترنت به راحتی از سر انگشتان ما قابل دسترسی است.

روند جستجو از طریق اینترنت یک کار سخت در پشت صحنه است.

بازیابی ویدیو مبتنی بر محتوا (CBVR) یکی از این روش های جستجو در اینترنت است. شناسایی حرکت انسان یکی از آنها در CBVR است. اگرچه تحقیقاتی در این زمینه ها وجود دارد، چالش ها نیز تا حدی حل شده اند. این مقاله همچنین به مسائل مربوط به شناسایی حرکت انسان با طراحی معماری به نام شبکه شناسایی حرکت انسانی فضایی-زمانی (STHARNet) می پردازد. سیستم پیشنهادی STHARNet در

یک سیستم CBVR یکپارچه شده است.عملکرد معماری پیشنهادی با آزمایش در سه مجموعه داده عمومی در دسترس: UCF Sports، KTH، و HMDB51 ارزیابی می شود. نتایج معماری پیشنهادی دلگرم کننده و بهتر از سایر روش های اخیر است.

1. **مقدمه:**

روش جستجو این امکان را برای انسان فراهم می کند که به راحتی به اطلاعات لازم دسترسی پیدا کند. از دستگاه های محلی شروع می شود و با استفاده از اینترنت در سطح جهانی رشد می کند. تاریخچه جستجو با فرهنگ لغت شروع می شود. به تدریج رشد می کند تا بتواند کتاب ها را با استفاده از ترتیب نمایه سازی حروف الفبا جستجو کند. هنگامی که دستگاه های الکترونیکی وارد زندگی انسان ها شد، تکنولوژی جستجو نیز وارد دستگاه های الکترونیکی شد . در دستگاه های محلی، جستجو به عنوان جستجو بر اساس متن و جستجو بر اساس صدا طبقه بندی شده است. در اینترنت، روش جستجو بیشتر به دو دسته طبقه بندی می شود: جستجو بر اساس تصویر و جستجو بر اساس ویدئو.جستجو بر اساس تصویر، بازیابی تصویر مبتنی بر محتوا (CBIR) نامیده می شود.جستجو بر اساس ویدیو، بازیابی ویدیو مبتنی بر محتوا نام دارد. در CBVR ورودی یک ویدیو یا صحنه کوتاه است که کاربر در جستجوی آن است. خروجی به همان تعدادیست که الگوریتم بازیابی می تواند آن محتوایی را که حاوی ویدیوی query داده شده است، را بازیابی کند. یک سیستم CBVR با چالش های متعددی مواجه است. یک سیستم CBVR باید به گونه ای طراحی شود که سیستم پیشنهادی قادر باشد ؛ تصاویر برش خورده، تصاویر فشرده و غیره را قبول کند. سیستم CBVR پیشنهادی یادگیری با نظارت را دنبال می کند. از این رو، سیستم به فازهای یادگیری و تست تقسیم می شود. که در هر دو فاز، ویدئو به صحنه ها و ویژگی های استخراج شده تقسیم می شوند. در فاز آموزش، ویژگی های استخراج شده در یک دیتاست ذخیره می شوند. در فاز تست، ویژگی استخراج شده با ویژگی دیتاست تطبیق داده است.

ویدیوهای حرکت انسان از ویژگی های مکانی و زمانی تشکیل شده است. ویژگی های مکانی آن هایی هستند که از یک فریم تنها استخراج شده اند. ویژگی های زمانی آنهایی هستند که از چندین فریم استخراج می شوند. آی تی حرکت را در ویدئو نشان می دهد. ویژگی های زمانی آنهایی هستند که CBVR را از CBIR متفاوت جلوه می دهند . این مقاله معماری HAR را توسعه داده و آن را با سیستم CBVR ادغام می کند. مشارکت های اصلی این مقاله عبارتند از:

* معماری شناسایی حرکت انسان فضایی-زمانی طراحی شده است
* ویژگی های مکانی از فریم های کلیدی استخراج می شوند.
* ویژگی های زمانی از تصویر انرژی حرکتی (MEI) ، سایر فریم ها استخراج می شوند.
* معماری HAR پیشنهادی در سیستم CBVR ادغام شده است

بقیه مقاله به این صورت سازماندهی شده است: بخش 2 برخی از روش های اخیر در سیستم های HAR و CBVR را مورد بحث قرار می دهد.بخش 3 سیستم STHAR پیشنهادی را با معماری و متودلوژِی (روش شناسی) آن شرح می دهد. همچنین ادغام STHAR در سیستم CBVR را نشان می دهد. بخش 4روش STHAR-CBVR پیشنهادی را با برخی آزمایش‌ها و تجزیه و تحلیل نشان می‌دهد.

بخش 5 کار را به پایان می رساند.

کارهای مرتبط:

+در این بخش برخی از روش‌های اخیر که برای مقایسه روش پیشنهادی استفاده می‌شوند، بحث می‌شود.

مدل های شبکه عصبی کانولوشنال (CNN) در بسیاری از سیستم های طبقه بندی استفاده می شود. CNN

می تواند هم به عنوان استخراج کننده و هم به عنوان طبقه بندی کننده استفاده شود. بسیاری از محققان CNN خود را می سازند

معماری برای HAR برخی از آن روش هایی که از CNN به عنوان مدل پایه خود استفاده می کنند به اختصار آورده شده است

در اینجا مورد بحث قرار گرفته است.

+یک روش انتخاب ویژگی به نام توزیع پواسون همراه با معیارهای تک متغیره

(PDaUM) [7] در سال 2021 توسعه یافت. تعدادی از ویژگی های CNN ذوب شده نامربوط هستند و

تعداد کمی از آنها اضافی هستند، که باعث پیش بینی نادرست در بین انسان های پیچیده می شود

اقدامات. این روش تنها قوی ترین ویژگی ها را انتخاب می کند که سپس به Extreme داده می شود

ماشین یادگیری (ELM) و Softmax برای تشخیص نهایی.

+معماری شبکه توجه جدید، به نام شبکه توجه چند سر آبشار

(CATNet) [21]، برای HAR توسعه یافته است. این نمایش‌های ویدیویی را با توجه دو سطحی ایجاد می‌کند، یعنی خودتوجهی محلی چند سر و توجه جهانی مبتنی بر رابطه.

با شروع از ویژگی های بخش تولید شده توسط شبکه های ستون فقرات، CATNet ابتدا یاد می گیرد

وزن توجه چندگانه برای هر بخش برای به تصویر کشیدن اهمیت ویژگی های محلی در یک

ابزارها و برنامه های چند رسانه ای

شیوه توجه به خود CATNet با وزن های توجه محلی، ویژگی های محلی را در آن ادغام می کند

چندین نمایش جهانی و سپس توجه سطح دوم را برای جهانی می آموزد

اطلاعات به صورت رابطه ای

+شبکه دیگری طراحی شده است که همبستگی های چندوجهی را بر روی دلخواه ثبت می کند

مهرهای زمانی [23]. این روش به عنوان یک شبکه تکمیلی و توسعه یافته روی یک چند وجهی عمل می کند

شبکه. جریان های مکانی و زمانی با استفاده از همجوشی شانون برای یادگیری یک آموزش از پیش آموزش دیده با هم ترکیب می شوند.

CNN. یک ویدیوی دوربرد ممکن است از همبستگی های مکانی-زمانی در زمان های دلخواه تشکیل شده باشد که

را می توان با تشکیل شبکه همبستگی از لایه های ساده و کاملاً متصل به دست آورد.

+یک توصیفگر مسیر جدید [22] بر اساس توصیفگرهای ویژگی محلی طراحی شده است

به عنوان هیستوگرام گرادیان های جهت دار (HOG)، هیستوگرام های جریان نوری (HOF) و حرکت

هیستوگرام مرزی همبستگی بین مسیرها و اقدام هدف را در بر می گیرد

توجه. یک نقشه برجسته بر اساس جریان نوری برای برجسته کردن مناطق معرفی شده است

از حرکت پیش زمینه ارتباط عمل مسیر و ارتباط عمل قاب به عنوان استفاده می شود

وزن ها برای شناسایی مسیرها و فریم های متمایز در یک ویدیو در طول رمزگذاری.

+ویژگی‌های مختلفی مانند انرژی، سینوسی، حرکات متمایز اعضای بدن و دکارتی سه بعدی

نمای شیب هموارسازی [5] برای استخراج شبح تمام بدن انسان استفاده شده است

ویژگی‌ها. ویژگی‌هایی که برای نشان دادن نقاط کلیدی وضعیت بدن انسان استخراج می‌شوند عبارتند از ظاهر دوبعدی غنی،

نقاط زاویه ای و خودهمبستگی چند نقطه ای. پس از استخراج نقاط کلیدی، یک مدل طبقه‌بندی سلسله مراتبی و بهینه‌سازی از طریق بهینه‌سازی پرتو و K-ary اعمال شده است.

الگوریتم هش درخت

+رایج ترین پلت فرم یادگیری بدون نظارت، PCANet، برای HAR استفاده می شود [2]. چندگانه

الگوهای تصویر انرژی حرکت کوتاه مدت (ST-MEI) برای استخراج زمانی محاسبه می‌شوند

امکانات. PCA-Net ویژگی های حرکتی سلسله مراتبی را ارائه می دهد که با استفاده از سفید کردن کاهش می یابد

PCA. به عنوان یک طبقه بندی، از ماشین بردار پشتیبانی (SVM) استفاده می شود.

+روش در [2] با ادغام ویژگی های مکانی و زمانی آموخته شده از a گسترش یافته است

PCANet در ترکیب با Bag-of-Features (BoF) و Vector of Locally Aggregated

طرح های رمزگذاری توصیفگرها (VLAD). در اینجا، یک طرح رمزگذاری برای نمایش اعمال می شود

ویژگی های PCANet فضایی-زمانی با ترکیب ویژگی ها

+یک الگوریتم موثر با استفاده از جریان نوری و تبدیل موجک توسعه داده شده است [20].

تبدیل دو برابری از طریق تبدیل موجک گابور (GWT) و Ridgelet انجام می شود.

تبدیل (RT). GWT یک بردار ویژگی را با محاسبه مقادیر آمار مرتبه اول تولید می کند

مقیاس ها و جهت گیری های مختلف یک ژست ورودی، که در برابر ترجمه، مقیاس بندی استحکام دارد،

و چرخش ویژگی‌های شکل وابسته به جهت‌گیری عملکرد انسان با استفاده از آن محاسبه می‌شود

RT. ادغام این ویژگی ها یک الگوریتم یکپارچه قوی به دست می دهد

+یک شبکه حافظه کوتاه مدت (LSTM) مبتنی بر توجه دو جریانی سرتاسری دارد

[4] برای تشخیص عمل توسعه داده شده است. این روش بر روی ویژگی های موثر تمرکز دارد

تصاویر ورودی اصلی و سطوح مختلف توجه به خروجی های هر ویژگی عمیق است

نقشه یک لایه همبستگی ویژگی عمیق برای تنظیم یادگیری عمیق پیاده سازی شده است

پارامترهای شبکه بر اساس قضاوت همبستگی

+یک شبکه باقیمانده عمیق فضایی-زمانی با توجه سلسله مراتبی (STDRN-HA)

برای HAR [10] طراحی شده است. در اولین لایه توجه، ویژگی ResNet کاملاً متصل است

ویژگی سریعتر R-CNN را برای ایجاد توجه مبتنی بر شی برای اشیاء هدف هدایت می کند. در

لایه توجه دوم، O-attention بیشتر ویژگی پیچیدگی ResNet را به سمت تسلیم هدایت می کند

توجه کل نگر (H-توجه). در لایه سوم توجه، نقشه های توجه از عمق استفاده می کنند

ویژگی هایی برای به دست آوردن ویژگی های افزایش دهنده توجه.

+در [18]، ویژگی‌های مکانی و زمانی با استفاده از یک سی‌ان‌ان سه بعدی پیشرفته استخراج می‌شوند.

نمونه‌هایی از یک ویدیو سپس این ویژگی‌ها به هم متصل می‌شوند تا نمایش‌های جهانی را تشکیل دهند.

ابزارها و برنامه های چند رسانه ای

که سپس برای آموزش یک SVM خطی برای طبقه بندی عمل استفاده می شود. A. Nadeem توسعه قوی

رویکرد ساختاری که ویژگی‌های چند بعدی را همراه با بازنمایی اعضای بدن کشف می‌کند

[11]. آنها از تجزیه و تحلیل متمایز درجه دوم با ویژگی های استخراج شده برای تخمین وضعیت انسان استفاده کردند

و تکنیک مارکوف حداکثر آنتروپی برای طبقه بندی.

+روش دیگری توسعه یافته است که تلاش می کند از ردیابی حرکت برای انسان استفاده کند

ردیابی و همچنین استخراج ویژگی های فضایی در یک توالی ویدیویی [6]. برای داشتن یک ویژگی قوی

بردار طبقه‌بندی، روش‌های مدل مخلوط گاوسی (GMM) و فیلتر کالمن (KF)

برای شناسایی و استخراج شخص متحرک و شبکه های عصبی بازگشتی دروازه ای استفاده می شود

برای جمع آوری ویژگی های هر فریم و پیش بینی عملکرد انسان استفاده می شود.

+چند مقاله تحقیقاتی اخیر مرتبط با سیستم های CBVR در زیر مورد بحث قرار می گیرد.

+یک سیستم CBVR برای یک مجموعه داده بزرگ طراحی شده است [15]. این روش از بردار استفاده می کند

امضاهای مبتنی بر حرکت برای توصیف محتوای بصری. همچنین از فریم های کلیدی برای بازیابی سریع استفاده می کند.

به جای استفاده از ویژگی های سطح پایین، ویژگی های سطح بالا در [12] در نظر گرفته شده است. این روش

از اشیاء به عنوان ویژگی استفاده می کند و شی را با تمام ویدیوهای دیگر در پایگاه داده مقایسه می کند.

+یک روش استخراج فریم کلیدی مبتنی بر اکشن در [14] توسعه یافته است. این

روش همچنین از فریم های کلیدی برای بازیابی سریع استفاده می کند. برای انتخاب فریم کلیدی، مناطق آموزنده

از هر فریم مشخص می شوند. چن و همکاران یک روش هش ویدیوی نظارت شده ایجاد کرده اند

بر اساس 3DCNN عمیق برای بازیابی ویدیو [3]. این روش ویژگی های ویدئویی را به

کدهای باینری با استفاده از تابع هش

+یک روش بازیابی ویدئویی بدون نظارت [24] که به طور همزمان توسعه یافته است

مدل سازی اطلاعات متنی درون فریم و درون فریم برای نمایش ویدئو با a

توپولوژی گراف که بر روی نقشه های مشخصه منطقه ای هرمی ساخته شده است. این روش

یک فریم را به یک زیرگراف منطقه ای هرمی تجزیه می کند و یک ویدیو را به یک منطقه ای تبدیل می کند

نمودار بنابراین، شبکه‌های کانولوشن گراف برای استخراج ویژگی‌هایی که ترکیب می‌شوند، استفاده شده‌اند

اطلاعات از انواع مختلف زمینه گروه هندسه بصری اصلاح شده (MVGG\_16)

[9] برای سیستم CBVR طراحی شده است. در این روش بازیابی فریم ویدئو می باشد

با اختصاص یک شاخص به تمام فایل های ویدئویی در پایگاه داده انجام می شود.

+یک CBVR چندوجهی [13] ساخته شده است که هم اطلاعات دیداری و هم اطلاعات صوتی را می گیرد

برای بازیابی ویدیوهای مرتبط برای کاربر در نظر گرفته شود. دو ماژول توسط این استفاده می شود

روشی برای مقابله با داده های صوتی و تصویری داده های ویدئویی برای شناسایی موارد مهم پردازش می شوند

فریم از عکس ها و توسط الگوریتم بهینه سازی شیر (LOA) به دست می آید. ویژگی ها هستند

از داده‌های بصری و با توجه به داده‌های صوتی، Mean Hilbert Envelope استخراج شده است

ویژگی های ضرایب (MHEC) و ضرایب پیش بینی خطی مغزی (LPCC) استخراج شده است. ویژگی های استخراج شده توسط الگوریتم هسته فازی C Mean (KFCM) خوشه بندی می شوند.

در نهایت، پایگاه داده ویژگی ها تشکیل می شود و در فرآیند تطبیق پرس و جو در طول دوره مورد استفاده قرار می گیرد

مرحله آزمایش

3 .روش HAR پیشنهادی:

سیستم HAR پیشنهادی با شبکه ای به نام STHAR طراحی شده است. معماری سیستم پیشنهادی در شکل 1 نشان داده شده است.

حرکت ها را می توان با استفاده از ویژگی‌های مکانی -زمانی، از یک ویدیو تشخیص داد. ویژگی های مکانی از فریم ها استخراج می شوند، در حالی که ویژگی های زمانی از اطلاعات حرکت استخراج می شوند. به جای استخراج ویژگی های مکانی از همه فریم ها، چند فریم نماینده از ویدیوی ورودی انتخاب می شوند. برای انتخاب فریم های نماینده یا فریم های کلیدی، روش گروه تطبیقی تصاویر (AGOP) [17] استفاده شده است.

شکل 1

در روش AGOP، ویدئو بر اساس برش صحنه به GOP ها (گروهی از تصاویر)تقسیم می شود. از هر GOP، اولین و آخرین فریم به عنوان فریم های کلیدی انتخاب می شوند. لازم به ذکر است که آخرین فریم از یک GOP اولین فریم از GOP بعدی است. تصویر انرژی حرکتی (MEI) از هر GOP برای ایجاد اطلاعات حرکتی برای آن GOP تولید می شود. ویژگی های زمانی، از MEI هر GOP استخراج شده است.هر دوی این ویژگی ها در شبکه STHAR ادغام و طبقه بندی شده اند.

3.1 شناسایی فریم کلیدی:

ویدئوی ورودی بر اساس حرکت اجسام به صحنه‌هایی (همچنین GOP نامیده شده است) در فریم ها تقسیم می‌شود.حرکت با مؤلفه همبستگی مشخص می شود. در این اثر ، از ضریب همبستگی پیرسون (PCC) استفاده شده است که به صورت زیر داده شده است:

فرمول 1

جایی که f و f p فریم هایی هستند که همبستگی برای آنها محاسبه می شود.f m و f m

p مقادیر میانگین فریم های f و f p به ترتیب هستند. مقدار مطلق PCC از 0 (بدون همبستگی) تا یک متغیر است (همبستگی کامل). تحلیل می شود که اگر مقدار PCC زیر 0.8 باشد، برش صحنه ای وجود ندارد. اگر از 0.8 بالاتر برود، یک برش وجود دارد. این مقدار توسط چند آزمایش انتخاب شده است.

یک ویدیوی نمونه برای این تحلیل گرفته شده است که صحنه های آن به صورت دستی شناسایی شده اند. شماره فریم های هر صحنه با روش بر پایه PCC بدست امده در مقایسه با تشخیص صحنه دستی برای مقادیر مختلف PCC در جدول 1 آورده شده است.

در جدول 1، برای صحنه 1، تعداد کل فریم ها با استفاده از تجسم دستی ، 10است. این

روش شناسایی صحنه مبتنی بر PCC تنها 1 فریم صحیح را برای صحنه 1 شناسایی می کند.فریم های باقی مانده به اشتباه به صحنه 2، صحنه 3 و غیره طبقه بندی می شوند. زمانی که مقدار آستانه (تریشلد)افزایش می یابد، تعداد فریم های صحیح در هر صحنه با تجسم دستی مطابقت دارد. زمانی که آستانه(تریشلد) 0.9 است، تعداد فریم های به درستی طبقه بندی شده، کاهش می یابد. در اینجا، فریم هایی که در

صحنه های بعدی به اشتباه در صحنه فعلی طبقه بندی می شوند (یعنی فریم ها در صحنه 2 هستند و به اشتباه به عنوان صحنه 1 طبقه بندی می شوند). از این رو، مشخص است که تعداد فریم ها در هر صحنه،زمانی که آستانه فریم کلیدی روی 0.8 تنظیم شده باشد، به درستی شناسایی می شود.

جدول 1

پس از برش صحنه، اولین فریم در هر GOP به عنوان فریم کلیدی انتخاب می شود. ویژگی های مکانی از این فریم های کلیدی استخراج شده است.

3.2 تصویر انرژی حرکت:

از هر GOP ایجاد شده در بخش فرعی قبلی، یک تصویر انرژی حرکتی تولید می شود

کدام ویژگی های زمانی استخراج می شوند. MEI یک تصویر حرکتی تجمعی باینری است که می باشد با استفاده از تفریق پس‌زمینه یا تفاوت فریم به دست می‌آید. تفاوت قاب یک است

راه حل ساده و کارآمد برای یک مشکل پس زمینه پویا. در این کار، چندین الگوی GOP MEI با استفاده از تفاوت قاب ساده برای حفظ حرکت محلی محاسبه می‌شوند.

اطلاعات و نمایش موثر حرکات انسانی که در تصویر رخ داده است

توالی.

فرض کنید که f xð Þ ; y; t و f ðx; y; t þ 1Þ نشان دهنده دو فریم متوالی در فیلم ورودی است

داده ها به ترتیب در زمان t و t + 1 عمل کردند. تفاوت مطلق فریم تفاوت xð Þ ; y; تی از اینها

دو فریم را می توان به صورت زیر بدست آورد:

فرمول 2

تفاوت (x، y، t) به صورت D(x، y، t) توسط آستانه Otsu دوتایی می شود. برای هر GOP، MEI به عنوان ایجاد می شود

فرمول 3

جایی که Egðx; y; tÞ تصویر انرژی حرکت در زمان t برای n تعداد فریم در یک GOP است.

شکل 2 فریم های متمایز فریم برای یک GOP و MEI بدست آمده را نشان می دهد.

کل ویدیوی اکشن (V) را می توان به صورت زیر بیان کرد

فرمول 4

جایی که G تعداد GOP است

شکل 2

3.3 شبکه ستاره:

یک معماری جدید برای استخراج ویژگی‌های مکانی و زمانی از فریم‌های کلیدی و

MEI که به ترتیب STHAR Net نامگذاری شده است. معماری شبکه STHAR است

در شکل 3 نشان داده شده است. این شبکه از ConvNets جریان فضایی و جریان زمانی تشکیل شده است.

ConvNets. ConvNets جریان فضایی بر روی فریم های کلیدی و ConvNet های جریان زمانی کار می کند

روی MEI کار کنید در این کار، شبکه آغازین با عادی سازی دسته ای (BN-Inception).

معماری به عنوان یک بلوک ساختمانی استفاده می شود، زیرا این شبکه تعادل خوبی بین کارایی دارد

و دقت نسبت انصراف برای ConvNet جریان فضایی 0.7 و برای موقت 0.8 تنظیم شده است.

ConvNet را استریم کنید. چارچوب ConvNet از 3 لایه تشکیل شده است. هر لایه از یک

نرمال سازی دسته ای، یک لایه پیچشی و یک فعال سازی واحد خطی اصلاح شده (ReLU)

تابع

3.4 بازیابی ویدئو مبتنی بر محتوا:

سیستم بازیابی ویدئو مبتنی بر محتوای مرسوم شامل مرحله آموزش و آزمایش است. که در

در مرحله آموزش، فیلم های آموزشی به شات های ویدیویی تقسیم می شوند که از آن ویژگی ها برخوردار است

استخراج و در یک پایگاه داده ذخیره می شود. در مرحله آزمایش، ویژگی‌های شات ویدیوی پرس و جو هستند

استخراج شده است. این ویژگی ها با تمام ویژگی های پایگاه داده ویژگی ها مطابقت دارند. چند تا بالا

عکس های ویدئویی مربوطه به عنوان خروجی بازیابی می شوند. در سیستم CBVR پیشنهادی، موارد فوق

مورد بحث STHAR یکپارچه شده است

شکل 3

شکل 4 سیستم پیشنهادی STHAR-CBVR را نشان می دهد. در سیستم STHAR، ویژگی های

برای به دست آوردن امتیاز هر عمل استفاده می شود. در سیستم CBVR، ویژگی ها ترکیب و ذخیره می شوند

در یک مجموعه داده ما ویژگی های مکانی و زمانی را با استفاده از یک پیچیدگی 1 بعدی به هم متصل می کنیم. برای

تصویر پرس و جو، ویژگی های مکانی-زمانی استخراج شده و با مجموعه داده ویژگی مقایسه می شود

با استفاده از معیار تشابه در این کار از فاصله اقلیدسی به عنوان معیار تشابه استفاده شده است.

مراحل ادغام STHAR و CBVR در زیر آورده شده است

+مرحله 1: فیلم ها را با استفاده از روش استخراج فریم کلیدی به صحنه ها تقسیم کنید.

+مرحله 2: ویژگی های مکانی و زمانی به طور جداگانه از فریم های کلیدی و

فریم های میانی به ترتیب.

+مرحله 3: نمرات مکانی و زمانی تولید شده و برای به دست آوردن نمره کلاس ترکیب می شوند

برای ویدیو

+مراحل 4: مراحل بالا سیستم STHAR را تشکیل می دهند

+مرحله 5: برای یک سیستم CBVR، ویژگی های مکانی و زمانی استخراج شده در مرحله 2 عبارتند از

ترکیب شده و در مجموعه داده های ویژگی ها ذخیره می شود.

+مرحله 6: برای یک صحنه پرس و جو معین، ویژگی های مکانی و زمانی با هم ترکیب شده و مطابقت دارند

با مجموعه داده های ویژگی ها با استفاده از اندازه گیری شباهت

+مرحله 7: تمام ویدیوهای مشابه که با ویژگی مطابقت دارند از پایگاه داده بازیابی می شوند.

شکل 4

الگوریتم سیستم پیشنهادی STHAR-CBVR در الگوریتم 1 نشان داده شده است.

الگوریتم 1 پیشنهاد سیستم STHAR-CBVR:

4 نتایج تجربی:

این بخش با ارائه نتایج به دست آمده، روش پیشنهادی را مورد بحث و تحلیل قرار می دهد. آی تی

همچنین شرح مختصری از مجموعه داده های مورد استفاده، مقایسه نتایج با روش های دیگر و

مطالعه فرسایش

4.1 شرح مجموعه داده:

روش پیشنهادی بر روی سه مجموعه داده در دسترس عموم آزمایش شده است: HMDB51 [8]، KTH [16]

و مجموعه داده UCF Sports [19]. جزئیات مجموعه داده در جدول 2 آورده شده است

سه مجموعه داده، HMDB51 اقدامات پیچیده ای دارد. مجموعه داده HMDB51 از 6766 ویدیو تشکیل شده است

از 51 اقدام در حالی که مجموعه داده ورزشی KTH و UCF به ترتیب شامل 6 و 10 اقدام است.

مجموعه داده به 80% آموزش و 20% تست تقسیم می شود. هر مجموعه داده دارای اندازه های مختلف است.

4.2 جزئیات پیاده سازی:

الگوریتم mini-batch Stochastic Gradient Descent (SGD) برای یادگیری شبکه استفاده می شود.

مولفه های. دو بهینه ساز SGD در این کار استفاده شده است (SGD1 و SGD2)

فرمول 5

جایی که λa یک اسکالر برای SGD1 است و λb یک اسکالر دیگر برای SGD2 برای متعادل کردن این دو است.

مشارکت دو بهینه ساز η نرخ یادگیری بهینه ساز ایجاد شده است. ηb است

نرخ یادگیری بهینه ساز SGD2. daand db دو افزایش SGD1 و SGD2 هستند

وزن آنها به ترتیب در هر تکرار. ما بهینه ساز لامبدا را به صورت پویا تغییر دادیم

از λa=1، λb=0 و λa=0، λb=1. η بر روی 1 تنظیم می شود که در هر 100 دوره در 0.1 ضرب می شود.

جدول 3 پارامترهای هایپر را برای هر دو شبکه نشان می دهد. برای افزایش داده ها،

چرخش افقی، برش گوشه و لرزش مقیاس استفاده می شود. روش پیشنهادی این است

تست شده بر روی دستگاه Intel Core i7 2.8 گیگاهرتز با پردازنده گرافیکی NVIDIA GeForce GTX 1050

4.3 معیارهای عملکرد:

روش پیشنهادی شامل دو کار تحقیقاتی عمده است: HAR و CBVR. عملکرد

سیستم HAR با استفاده از دقت اندازه گیری می شود. به طور مشابه، عملکرد CBVR

سیستم با استفاده از دقت، یادآوری، امتیاز F و حساسیت اندازه گیری می شود. تمام معیارهای عملکرد

در این کار از 4 مقدار اصلی استفاده شده است: مثبت واقعی (TP)، منفی واقعی (TN)، مثبت کاذب

(FP) و منفی کاذب (FN). فرمول تمام معیارها در جدول 4 نشان داده شده است.

4.4 نتایج:

از آنجایی که در این تحقیق از دو فاز اصلی استفاده می شود، هر دو فاز بر حسب مورد تجزیه و تحلیل می شوند

کارایی. نتایج بصری و کمی در این بخش مورد تجزیه و تحلیل قرار می گیرد. دقت

CBVR بر اساس ارتباط مجموعه خروجی تصاویر یا ویدیوها با یک جستجوی خاص است.

یک کاربر سطح کیفیت را برای ارزیابی CBVR تعریف می کند. دقت این سیستم ها

صرفاً ذهنی است که نیاز به مداخله انسانی در طول ارزیابی دارد.

جدول 2

جدول 3

جدول 4

سبه یک موضوع به طور کلی یک پرس و جو نشان داده می شود و از آن خواسته می شود تا تصویر یا ویدیوی حاصل را رتبه بندی کند

بخش ها در مقیاس به عنوان مثال، در یک پایگاه داده ویدیویی ممکن است از کاربر خواسته شود که به آن امتیاز دهد

کیفیت انتخاب در مقیاسی از "ارتباط بالا" (رتبه - 1) تا "ارتباط کم"

(رتبه - 10). ما 10 نتیجه مرتبط برتر تصویر پرس و جو را در نظر می گیریم.

دقت با استفاده از 10 خروجی برتر نتایج بازیابی شده و 10 نتیجه برتر به دست می آید.

با ارزیابی ذهنی به دست می آید. نتایج بصری به دست آمده توسط سیستم STHAR-CBVR عبارتند از

در جدول 5 نشان داده شده است. تنها 5 فریم از نتایج بازیابی شده در جدول نشان داده شده است. همه

فریم های جدول با مجموعه داده HMDB51 مطابقت دارد. از آنجایی که این تحقیق ادغام HAR است

و CBVR، نام کلاس نیز در جدول 5 نشان داده شده است

جدول 5

جدول 6

تفاوت HAR و CBVR در این است که HAR نام عمل را نشان می دهد

کلاس؛ در حالی که CBVR ویدیو را بر اساس ویدیوی پرس و جو نشان می دهد. سیستم پیشنهادی STHAR CBVR همچنین برخی ویدیوهای دیگر مرتبط با محتوا را نشان می دهد. اما فقط چند فریم هستند

در جدول 5 نشان داده شده است.

از جدول 5 مشاهده می شود که خروجی های بازیابی شده بسیار شبیه به چارچوب پرس و جو هستند.

طبقاتی که در جدول 5 نشان داده شده اند، خروجی سیستم HAR پیشنهادی هستند. بنابراین

STHAR-CBVR پیشنهادی قادر به شناسایی کلاس و به دست آوردن ویدیوهای مشابه از مجموعه داده برای است

صحنه پرس و جو داده شده دقت به دست آمده توسط سیستم STHAR پیشنهادی در نشان داده شده است

جدول 6.

از جدول 6 مشاهده می شود که دقت به دست آمده توسط سیستم STHAR پیشنهادی برابر است

بیش از 90٪ برای همه مجموعه داده ها. همچنین مطالعه شده است که دقت، یادآوری، امتیاز F،

حساسیت و دقت سیستم پیشنهادی STHAR-CBVR در محدوده 71-75 درصد است. این

دقت سیستم بازیابی بسیار کمتر از سیستم تشخیص است. سردرگمی

ماتریس های سیستم STHAR پیشنهادی برای مجموعه داده های KTH و UCF Sports Action داده شده است

در شکل 5.

در مجموعه داده KTH، برای 4 کلاس به دقت تقریباً 99٪ رسیدیم. در اقدام ورزشی UCF

مجموعه داده، ما به حداکثر دقت برای کلاس های Diving-side، Golf-swing و Swing-Bench رسیدیم.

سیستم STHAR پیشنهادی به دقت 99 درصد در 4 کلاس از مجموعه داده های ورزشی UCF دست می یابد

4.5 مقایسه روش های پیشنهادی با روش های اخیر:

روش های پیشنهادی STHAR و STHAR-CBVR به طور جداگانه با روش های اخیر مقایسه می شوند

روش های [1، 2، 4-7، 10، 11، 14، 15، 18، 20-23] که در بخش 2 مورد بحث قرار گرفته اند. جدول 7 نشان می دهد

مقایسه روش پیشنهادی STHAR با روش های اخیر برای مجموعه داده های آزمایش شده

از جدول 7، مطالعه شده است که 0.7٪ افزایش در دقت در مجموعه داده HMDB51 وجود دارد.

در مقایسه با مدل چوب شبه دو بعدی [5]. به طور مشابه، 0.4٪ افزایش در دقت وجود دارد

در مجموعه داده KTH. اما در مجموعه داده UCF Sports، دقت 0.2٪ کاهش می یابد. جدول 8 نشان می دهد

مقایسه سیستم بازیابی با سایر سیستم های بازیابی

از جدول 8 استنباط می شود که روش پیشنهادی دقت بالاتری نسبت به اخیر دارد

روش‌ها در مجموعه داده‌های HMDB51 و KTH.

4.6 مطالعه ابلیشن:

روش پیشنهادی با دو آزمایش مختلف ارزیابی می‌شود. فضایی و زمانی

ویژگی ها با استفاده از آنها به طور جداگانه مورد مطالعه قرار می گیرند. مدل های شبکه از پیش آموزش دیده نیز متنوع هستند

برای اندازه گیری دقت روش های پیشنهادی. جدول 9 اولین مطالعه و جدول 10 را نشان می دهد

مطالعه دوم را نشان می دهد

شکل 5

جدول 7

از جدول فوق، بررسی می شود که شبکه دو جریانی به دقت بالاتری دست می یابد

نسبت به جریان های منفرد در همه مجموعه داده ها. دقت زمانی که فقط فضایی است بسیار کمتر است

ConvNet ها استفاده می شوند. محبوب ترین Googlenet و VGG16 برای فضایی و زمانی استفاده می شود

ConvNets و میانگین نتایج در جدول 10 نشان داده شده است.

از جدول 10، مشخص شده است که مدل شبکه BN-Inception به دقت بالاتری دست می یابد

مدل های GoogleNet و VGG16

جدول 8

جدول 9

جدول 10

5 نتیجه گیری:

سیستم‌های بازیابی ویدیوی مبتنی بر محتوا و تشخیص کنش انسانی، مناطق وسیعی را اشغال می‌کنند

حوزه پژوهشی. این مقاله با معرفی یک STHARNet جدید، هر دو روش را ترکیب می‌کند

سیستم HAR. سیستم پیشنهادی با سیستم CBVR یکپارچه شده است و هر دو را برآورده می کند

سیستم های. این بر روی مجموعه داده های اصلی مانند مجموعه داده های عملی KTH، HMDB51 و UCF Sports آزمایش شده است

و کارایی خود را نسبت به سایر روش ها ثابت کرد. STHAR و STHAR-CBVR پیشنهادی

سیستم ها به ترتیب به طور متوسط 98٪ و 74٪ برای همه مجموعه داده ها دقت دارند.

سیستم پیشنهادی را می توان برای استفاده از برخی دیگر از شبکه های از پیش آموزش دیده و می توان گسترش داد

روی مجموعه داده های پیچیده تر آزمایش شود.