بازنمایی ویدیو

امیر حسین بختیاری

چکیده

برای طبقه بندی ویدیو ها در بازشناسی حرکت به یک بردار بازنمایی ویدیو نیاز داریم. بردار بازنمایی معمولا با استفاده از ویژگی های حرکتی و ظاهری هر قاب از ویدیو به دست می آیند. برای رسیدن به بردار بازنمایی با استفاده از ویژگی های قاب از روش های مختلف ادغام استفاده می شود. در این نوشتار دو روش جدید برای بازنمایی ویدیو معرفی می کنیم. در هر دو روش دنباله به دست آمده از هر قاب را به صورت n سری زمانی در نظر می گیریم. تفاوت اصلی بین این دو روش از این جا به بعد است. از روش اول بیشتر برای ویدیو های اول شخص و از روش دوم برای ویدیو های سوم شخص استفاده می کنیم. روش دوم مبتنی بر کانولوشن یک بعدی است. در پایان با مقایسه نتایج به دست آمده از هر یک، با روش های متداول پیشین بهبود عملکرد را به صورت کاملا محسوس مشاهده می کنیم.

کلیدواژه‌ها

بازنمایی ویدیو، ویژگی های ظاهری، ویژگی های حرکتی، ادغام زمانی، کانولوشن یک بعدی، ویدیوی اول شخص

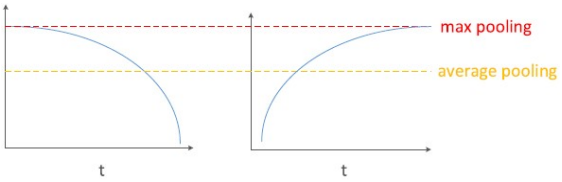
۱ مقدمه

یک ویدیو از دنباله ای از قاب[1] ها تشکیل شده است. این دنباله نمایش دهنده تغییر محتوای ویدیو است. هر قاب دارای ویژگی[2] هایی است که به دو دسته ظاهری[3] و حرکتی[4] تقسیم می شوند. بنابراین می توان هر ویدیو را به صورت دنباله ای از ویژگی های قاب (ظاهری یا حرکتی) نشان داد. هدف در این مقاله آن است که بتوان به یک بازنمایی[5] برداری برای هر ویدیو دست یافت. از این بازنمایش برداری می توان برای مقاصد گوناگون، از جمله بازشناسی حرکت[6] استفاده نمود. یک روش متداول برای رسیدن به چنین نمایشی استفاده از ویژگی های ظاهری یا حرکتی قاب ها است. برای تبدیل دنباله ویژگی های قاب یک ویدیو به یک بازنمایش برداری، می توان از ادغام زمانی[7] بهره جست.

در این نوشتار دو روش برای ادغام زمانی توضیح داده شده است. روش اول برای ویدیو های اول شخص است، و روش دوم بیشتر برای ویدیو های سوم شخص مورد استفاده قرار می گیرد. در ویدیوی اول شخص، فردی که دوربین را پوشیده است (چنین دوربینی معمولا بخشی از یک ابزارک[8] پوشیدنی است) فعالانه درگیر رویداد هایی می شود که در حال ضبط شدن است؛ در حالی که در ویدیوی سوم شخص چنین نیست. ادغام های زمانی متداول عبارتند از ادغام ماکزیمم و ادغام میانگین. یکی از مشکلات چنین روش هایی در نظر نگرفتن ترتیب قاب ها است. نادیده گرفتن چنین ترتیبی منجر به از دست رفتن اطلاعات حرکتی می شود. برای مثال می توان به زوج حرکت نشستن و برخاستن اشاره نمود. طبعا با روش های متداول ادغام زمانی و نادیده گرفتن ترتیب قاب ها، این دو حرکت قابل تشخیص از یکدیگر نخواهند بود، چنانچه در شکل ۱ مشاهده می شود. برای دریافت بهتر اطلاعات ترتیبی

[1] Frame
[2] Features
[3] Appearance
[4] Motion
[5] Representation
[6] Action Recognition
[7] Temporal Pooling
[8] Gadget

قاب ها، دو روش جدید برای ادغام زمانی ارائه می شود. روش دوم از شبکه های عصبی کانولوشنی الهام گرفته شده و به صورت ویژه از لایه های کانولوشنی یک بعدی استفاده می کند.



شکل ۱ نمایش ادغام ماکزیمم و میانگین برای دو سیگنال یک بعدی

۲ ادغام زمانی به روش اول (PoT[1])

برای بازنمایی هر ویدیو روال کار با استخراج قاب های آن آغاز می شود. پس از آن، ویژگی های هر قاب اعم از ظاهری یا حرکتی را به دست می آوریم. در ادامه ادغام زمانی بر روی این دنباله از ویژگی ها انجام می شود؛ و در پایان یک بازنمایی برداری برای ویدیو خواهیم داشت. در نهایت می توان این بازنمایی را به یک طبقه بند[2] داد تا به عنوان مثال نوع حرکت تشخیص داده شود.

روند کلی در این روش در ادامه توضیح داده می شود. ابتدا ویژگی های ظاهری/حرکتی از هر قاب استخراج می شود. در نتیجه یک دنباله از بردار ویژگی های n بعدی به دست می آید. n اندازه بردار به دست آمده از هر قاب است. در این روش دنباله به دست آمده به صورت n عدد سری زمانی در نظر گرفته می شود. ایده کلی آن است که تغییرات هر یک از n عنصر موجود در بردار ویژگی، در طول زمان زیر نظر گرفته شود. اگر تعداد قاب های یک ویدیو را m در نظر بگیریم؛ تغییرات عنصر n ام در هر یک از m بردار ویژگی، یکی سری زمانی به طول m خواهد بود. در مرحله بعدی ادغام زمانی اعمال می شود. برای این منظور مجموعه ای از فیلتر های زمانی (به عنوان مثال بازه های زمانی) به هر یک از سری های زمانی اعمال می شود. سپس در هر یک از این بازه های زمانی چندین عمل برای ادغام (برای مثال ادغام ماکزیمم، میانگین، گرادیان ها ...) انجام می شود. در نهایت، برای به دست آوردن بازنمایی ویدیو، نتایج ادغام به یکدیگر چسبانده می شوند.

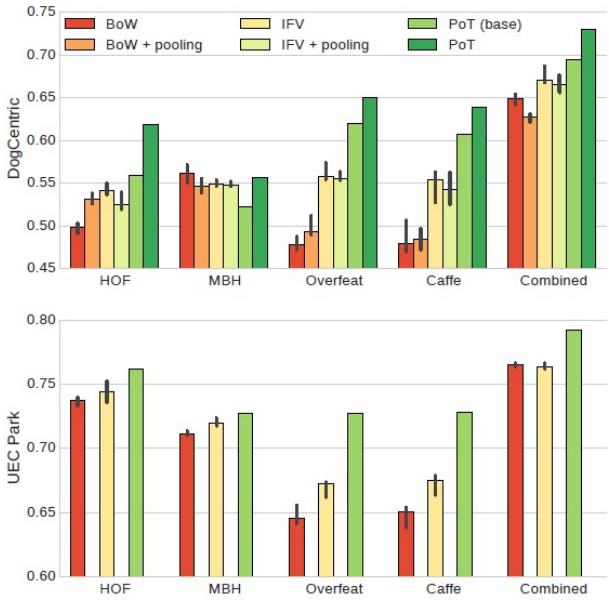
در این روش، برای به دست آوردن ویژگی های ظاهری قاب از دو شبکه CNN[3] استفاده می شود. این شبکه ها از پیش بر روی مجموعه داده ImageNet آموزش دیده اند. ویژگی های به دست آمده از هر یک از این دو شبکه ۴۰۹۶ بعدی هستند. برای ویژگی های حرکتی نیز از HOF[4] و MBH[5] استفاده می شود که به ترتیب ۲۰۰ و ۴۰۰ بعدی می باشند. ویژگی های حرکتی یاد شده مبتنی بر جریان بصری[6] هستند.

2-1عملگر های ادغام زمانی

در روش گفته شده که موسوم به PoT است؛ چهار نوع ادغام زمانی در هر یک از فیلترها (بازه ها) ی زمانی انجام می شود، ادغام ماکزیمم، ادغام مجموع و دو نوع ادغام جدید دیگر با نام هیستوگرام گرادیان سری زمانی[7]. برای اولین ادغام جدید، تعداد گرادیان های مثبت (و منفی) در هر بازه زمانی شمرده می شود. برای ادغام نوع دوم، مجموع مقادیر گرادیان های مثبت (یا منفی) در هر بازه محاسبه می شود.

2-2آزمایش روش پیشنهادی

این روش بر روی دو مجموعه داده با نام های DogCentric و UEC Park مورد بررسی قرار گرفت. بهبود عملکرد به صورت قابل توجه ای نسبت به روش های پیشین مشهود است. نتایج به دست آمده در شکل ۲ قابل مشاهده است.



شکل ۲ دقت کلاس بندی روش های مختلف بازنمایی ویدیو

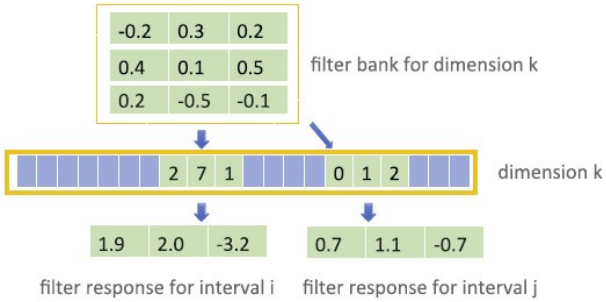
۳ ادغام زمانی به روش دوم (Order-aware)

روش دوم به صورت ویژه ای برای ویدیو های سوم شخص مناسب است. در این روش اصول کلی شباهت زیادی به روش اول دارند. ابتدا این شباهت ذکر شده، سپس به بیان تفاوت ها می پردازیم. چنانچه پیش از این گفته شد؛ پیش از هر چیز قاب ها

[1] Pooled Time Series
[2] Classifier
[3] Convolutional Neural Networks
[4] Histogram of Optical Flows
[5] Motion Boundary Histogram
[6] Optical Flow
[7] Histogram of Time Series Gradients

استخراج می شوند. سپس ویژگی های ظاهری و حرکتی آنها را به دست می آوریم. ادغام زمانی صورت می گیرد و در نهایت نتایج به یکدیگر چسبانده می شوند. در این روش نیز چنانچه اندازه بردار ویژگی به دست آمده از هر قاب را n فرض کنیم؛ پس از استخراج ویژگی های هر قاب، n سری زمانی خواهیم داشت.

هر سری زمانی به عنوان یک سیگنال یک بعدی در نظر گرفته می شود. در این جا اولین تفاوت آشکار می شود. به هر سیگنال یک بعدی، یک بانک فیلتر منحصر به فرد اعمال می شود. پاسخ به دست آمده نتیجه کانولوشن یک بعدی فیلتر، در سری زمانی است. چگونگی اجرای عمل کانولوشن در شکل ۳ به صورت مشخص نشان داده شده است.

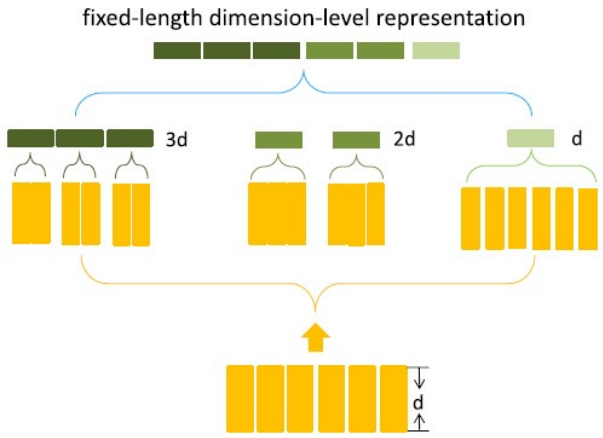


شکل ۳ نمایش کانولوشن یک بعدی. در این شکل تعداد فیلتر های موجود در بانک فیلتر، ۳ و اندازه بازه زمانی نیز، ۳ در نظر گرفته شده است.

در مرحله بعد، پاسخ های فیلتر به دست آمده از مرحله قبل را، به m_i قسمت افراز می کنیم. اندیس i شان دهنده مرحله افراز است. در روش پیشنهادی دو مرحله برای افراز وجود دارد. به عنوان مثال در مرحله اول، پاسخ های فیلتر سیگنال n ام، به دو قسمت و در مرحله دوم به ۳ قسمت افراز می شود. در هر مرحله از افراز، برای هر یک از قسمت ها، ادغام زمانی ماکزیمم صورت می گیرد. در پایان نتایج به دست آمده برای هر قسمت در هر مرحله از ادغام، به یکدیگر چسبانده می شوند. این عمل برای هر یک از n سیگنال یک بعدی انجام می شود. بازنمایی نهایی ویدیو، از به هم چسباندن نتایج حاصل از هر یک از سیگنال های یک بعدی به دست می آید. مراحل افراز و ادغام در شکل ۴ نمایش داده شده است.

یکی از مشکلات بازنمایی ویدیو با استفاده از کانولوشن، مساوی نبودن تعداد قاب های هر ویدیو است. یکسان نبودن تعداد قاب ها، منجر به بردار بازنمایی ویدیو با طول های مختلف می شود. از آنجا که این بردار ها قرار است به دست بند داده شوند، می توانند مشکلاتی را به وجود آورند. با استفاده از روش افراز و ادغام پیشنهادی این مشکل نیز برطرف می شود. می توان دید که صرف نظر از تعداد قاب ها، همواره بردار بازنمایی با طول ثابت به دست می آید.

در این روش برای استخراج ویژگی های ظاهری از یک شبکه CNN با نام AlexNet استفاده می شود که پیشتر بر روی مجموعه داده ImageNet آموزش دیده شده است. برای به دست آوردن ویژگی های حرکتی هر قاب هم از استفاده IDT[1] استفاده می شود.



شکل ٤ نمایش ادغام هرمی زمانی

روش گفته شده بر روی سه مجموعه داده HMDB51، UCF101 و Hollywood2 مورد بررسی قرار گرفت و نتایج به دست آمده، نشان از بهبود عملکرد، نسبت به روش های پیشین دارد. می توان مقایسه های صورت گرفته را در جدول ۱ و جدول ۲ مشاهده نمود.

جدول ۱

Comparison of the proposed pooling method to the baselines on HMDB51 using appearance information or motion information.

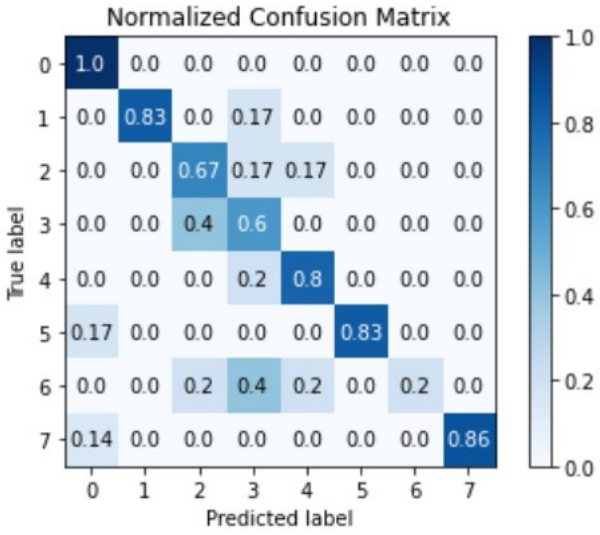
Appearance	AP	37.5%
	MP	36.5%
	PoT (no TP) [46]	36.5%
	TP	39.2%
	Ours (MP)	**40.8%**
	Ours (TP)	**41.6%**
Motion	AP	50.9%
	MP	50.6%
	TP	54.7%
	Ours (MP)	52.8%
	Ours (TP)	**55.0%**

[1] Improved Dense Trajectory

جدول ۲

Comparison of the proposed pooling method to the baselines on UCF101 using appearance information or motion information.

Appearance	AP	66.3%
	MP	67.4%
	PoT (no TP) [46]	67.5%
	TP	68.5%
	Ours (MP)	**69.3%**
	Ours (TP)	**70.4%**
Motion	AP	80.0%
	MP	80.2%
	TP	81.6%
	Ours (MP)	81.0%
	Ours (TP)	**82.1%**



شکل ۵ ماتریس در همی برای روش **1D CNN**

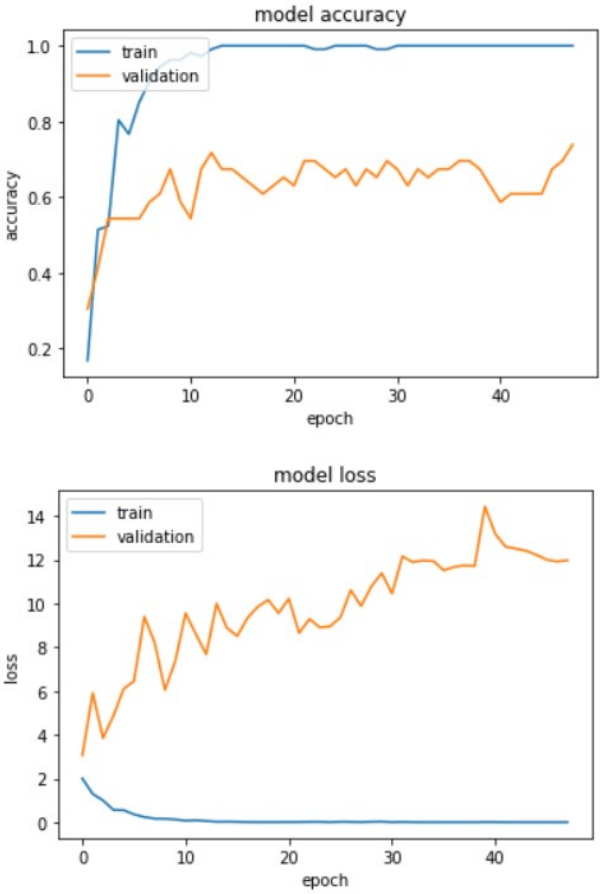
۴ پیاده سازی

برای پیاده سازی از روش دوم الهام گرفته شد. اما مجموعه داده و جزییات پیاده سازی با روش گفته شده متفاوت است. برای مجموعه داده از NUSFPID استفاده شده است. این مجموعه داده، مجموعه کوچکی از ویدیو های اول شخص است.

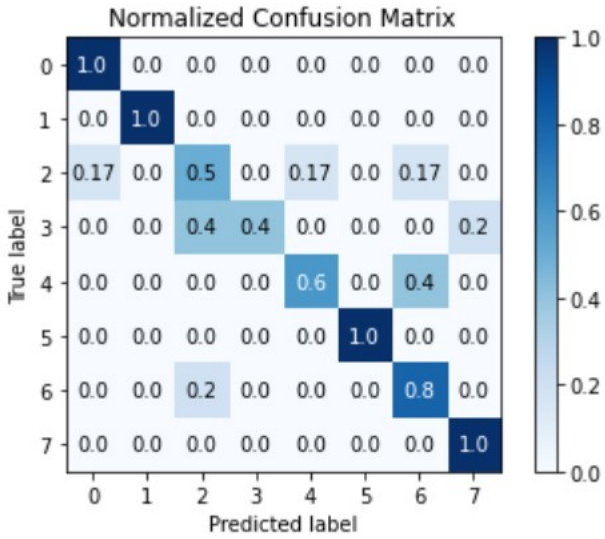
از آنجا که تعداد قاب های هر ویدیو در این مجموعه متفاوت است؛ قاب ها متناسب با تعداد کل آنها، نمونه برداری شده اند. به این ترتیب که کمترین تعداد قاب ۵۴ است. بنابر این اگر تعداد قاب ویدیویی بین ۵۴ تا کمتر از دو برابر آن باشد، ۵۴ قاب اول انتخاب می شوند. و در نهایت اگر تعداد قاب های یک ویدیو، بین ۸ تا ۹ برابر کمترین مقدار (۵۴) باشند، از هر ۸ قاب یکی انتخاب می شود. بدین ترتیب مشکل مساوی نبودن تعداد قاب ها برطرف می شود.

در این روش تنها ویژگی های ظاهری هر فریم استخراج شده است. برای استخراج ویژگی از شبکه vgg16 آموزش دیده بر روی مجموعه داده ImageNet استفاده می شود. ویژگی های ظاهری هر قاب از لایه FC2 گرفته می شود. دنباله ویژگی ها به صورت یک سری زمانی با طول ۵۴ در نظر گرفته می شوند. اندازه هر گام زمانی ۴۰۹۶ (اندازه بردار ویژگی هر قاب) است. یکی شبکه کانولوشنی یک بعدی این سری زمانی را گرفته و خروجی آن را به یک طبقه بند می دهد.

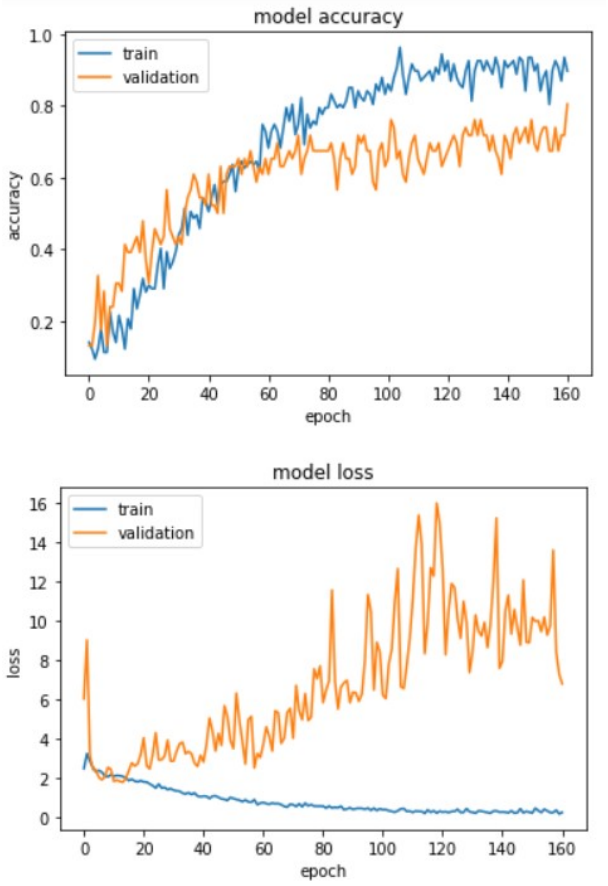
نتایج به دست آمده در مقایسه با روش ادغام ماکزیمم چندان رضایت بخش نیست. یک علت می تواند کوچک بودن مجموعه داده باشد. این کوچک بودن منجر به بیش برازش شبکه می شود و در نتیجه شبکه به خوبی آموزش نمی بیند. نتایج برای کانولوشن یک بعدی و ادغام ماکزیمم نمایش داده شده است.



شکل ۶ نمودار دقت و خطا برای روش **1D CNN**



شکل ۷ ماتریس درهمی برای روش ادغام ماکزیمم



شکل ۸ نمودار دقت و خطا برای روش ادغام ماکزیمم

مراجع

[1] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled Motion Features for First-Person Videos," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, no. Figure 1, pp. 896–904, Oct. 2015, doi: 10.1109/CVPR.2015.7298691.

[2] P. Wang, L. Liu, C. Shen, and H. T. Shen, "Order-aware convolutional pooling for video based action recognition," *Pattern Recognit.*, vol. 91, pp. 357–365, 2019, doi: 10.1016/j.patcog.2019.03.002.