تشخیص وقایع بصری به‌کمک اطلاعات مکانی-زمانی سیگنال ویدئو

محمد سلطانیان[1] و شاهرخ قائم‌مقامی*[2]

[2 و1] دانشکده مهندسی برق و پژوهشکده الکترونیک، دانشگاه صنعتی شریف، تهران، ایران

[1] گروه علوم کامپیوتر، دانشکده علوم ریاضی و کامپیوتر، دانشگاه خوارزمی، تهران، ایران

چکیده

در این مقاله، تشخیص وقایع بصری در ویدئو، با بهره‌گیری از اطلاعات زمانی سیگنال، به‌صورت تحلیلی موردتوجه قرار دارد. با استفاده از یادگیری انتقالی، توصیف‌گرهای آموزش‌دیده روی تصاویر به ویدئو اعمال می‌شوند تا تشخیص وقایع را با استفاده از منابع محاسباتی محدود، ممکن سازند. در این مقاله، یک شبکه عصبی کانولوشنی به‌عنوان استخراج‌کننده نمرات مفاهیم از قاب‌های ویدئو به‌کار می‌رود. ابتدا پارامترهای این شبکه روی زیرمجموعه‌ای از داده‌های آموزش تنظیم دقیق می‌شوند؛ سپس، توصیف‌گرهای خروجی از لایه‌های تمام‌متصل آن به‌عنوان توصیف‌گر سطح قاب مورداستفاده قرار می‌گیرند. توصیف‌گرهای به‌دست‌آمده، کدگذاری و درنهایت نرمالیزه‌سازی و طبقه‌بندی می‌شوند. نوآوری عمده این مقاله، ترکیب اطلاعات زمانی ویدئو در کدگذاری توصیف‌گرهای آن است. گنجاندن ساختاری اطلاعات بصری در فرایند کدگذاری توصیف‌گرهای ویدئویی،، اغلب نادیده گرفته می‌شود. این موضوع به کاهش دقت منجر می‌شود. برای حل این مسأله، یک روش کدگذاری نوین ارائه می‌شود که مصالحه بین پیچیدگی محاسبات و دقت در شناسایی وقایع ویدیویی را بهبود می‌دهد. در این کدگذاری، بعد زمانی سیگنال ویدئویی برای ساخت یک بردار مکانی-زمانی از توصیف‌گرهای مجتمع محلی (VLAD) استفاده، سپس نشان داده می‌شود که کدگذاری پیشنهادی ماهیتاً یک مسأله بهینه‌سازی است که با الگوریتم‌های موجود به‌راحتی قابل‌حل است. در مقایسه با بهترین روش‌های موجود در حوزه تشخیص وقایع بصری مبتنی بر توصیف‌گرهای سطح قاب، روش پیشنهادی مدل بهتری را از ویدئو ارائه می‌کند. روش ارائه‌شده بر حسب سه معیار میانگین دقت متوسط، میانگین فراخوانی متوسط و معیار F به عملکرد بالاتری بر روی هر دو مجموعه‌دادهٔ آزمون مورد بررسی دست می‌یابد. نتایج به‌دست‌آمده توانمندی روش پیشنهادی را در بهبود عملکرد سامانه‌های تشخیص وقایع بصری تأیید می‌کنند.

واژگان کلیدی: شبکه عصبی کانولوشنی، ادغام میانگین، ادغام بیشینه، ماشین بردار پشتیبان، بردار توصیف‌گرهای مجتمع محلی

Recognition of Visual Events using Spatio-Temporal Information of the Video Signal

Mohammad Soltanian[1] & Shahrokh Ghaemmaghami*[2]

[1,2] Department of Electrical Engineering and Electronics Research Institute
Sharif University of Technology, Tehran, Iran

[1] Department of Computer Sciences, Faculty of Mathematical Sciences and Computer,
Kharazmi University, Tehran, Iran

Abstract

Recognition of visual events as a video analysis task has become popular in machine learning community. While the traditional approaches for detection of video events have been used for a long time, the recently evolved deep learning based methods have revolutionized this area. They have enabled event recognition systems to achieve detection rates which were not reachable by traditional approaches. Convolutional neural networks (CNNs) are among the most popular types of deep networks utilized in both imaga and video recognition tasks. They are initially made up of several convolutional layers, each

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

•تاریخ ارسال مقاله: ۱۳۹۷/۰۸/۲۱ •تاریخ پذیرش: ۱۳۹۷/۱۱/۳۰ •تاریخ انتشار: ۱۴۰۰/۰۳/۰۱ •نوع مطالعه: بنیادی

[DOI: 10.52547/jsdp.18.1.134]

of which followed by proper activation and possibly pooling layers. They often encompass one or more fully connected layers as the last layers. The favorite property of them in this work is the ability of CNNs to extract mid-level features from video frames. Actually, despite traditional approaches based on low-level visual features, the CNNs make it possible to extract higher level semantic features from the video frames.

The focus of this paper is on recognition of visual events in video using CNNs. In this work, image trained descriptors are used to make video recognition can be done with low computational complexity. A tuned CNN is used as the frame descriptor and its fully connected layers are utilized as concept detectors. So, the featue maps of activation layers following fully connected layers act as feature vectors. These feature vectors (concept vectors) are actually the mid-level features which are a better video representation than the low level features. The obtained mid-level features can partially fill the semantic gap between low level features and high level semantics of video.

The obtained descriptors from the CNNs for each video are varying length stack of feature vectors. To make the obtained descriptors organized and prepared for clasification, they must be properly encoded. The coded descriptors are then normalized and classified. The normaliztion may consist of conventional $\ell_1$and ℓ_2 normalization or more advanced power-law normalization. The main purpose of normalization is to change the distribution of descriptor values in a way to make them more uniformly distributed. So, very large or very small descriptors could have a more balanced impact on recognition of events.

The main novelty of this paper is that spatial and temporal information in mid-level features are employed to construct a suitable coding procedure. We use temporal information in coding of video descriptors. Such information is often ignored, resulting in reduced coding efficiency. Hence, a new coding is proposed which improves the trade-off between the computation complexity of the recognition scheme and the accuracy in identifying video events.

It is also shown that the proposed coding is in the form of an optimization problem that can be solved with existing algorithms. The optimization problem is initially non-convex and not solvable with the existing methods in polynomial time. So, it is transformed to a convex form which makes it a well defined optimization problem. While there are many methods to handle these types of convex optimization problems, we chose to use a strong convex optimization library to efficiently solve the problem and obtain the video descriptors.

To confirm the effectiveness of the proposed descriptor coding method, extensive experiments are done on two large public datasets: Columbia consumer video (CCV) dataset and ActivityNet dataset. Both CCV and ActivityNet are popular publically available video event recognition datasets, with standard train/test splits, which are large enough to be used as reasonable benchmarks in video recognition tasks.

Compared to the best practices available in the field of detecting visual events, the proposed method provides a better model of video and a much better mean average precision, mean average recall, and F score on the test set of CCV and ActivityNet datasets. The presented method not only improves the performance in terms of accuracy, but also reduces the computational cost with respect to those of the state of the art. The experiments vividly confirm the potential of the proposed method in improving the performance of visual recognition systems, especially in supervised video event detection.

Keywords: Convolutional neural network, Average pooling, Max pooling, Support vector machine, Vector of locally aggregated descriptors.

۱- مقدمه

تحلیل معنایی ویدئو، زمینه‌ای است که در سال‌های اخیر توجه زیادی را به خود جلب کرده است. یکی از دلایل این توجه، نیاز به ساخت سامانه‌های هوشمندی است که بتوانند با استفاده از یک دوربین با انسان تعامل داشته باشند یا وظایفی همچون نظارت ویدئویی، خلاصه‌سازی ویدئو، و یا جستجوی مفهومی در مجموعه عظیمی از ویدئوها را انجام دهند. به‌طور خاص، تحلیل معنایی ویدئو با استفاده از وقایع بصری، به آشکارسازی و بازشناسی وقایع بصری در ویدئو می‌پردازد. با استفاده از این تحلیل می‌توان به مدل واقع‌گرایانه‌تری از سیگنال ویدئو، که با ادراک انسانی تطابق بیشتری دارد، دست‌یافت. آشکارسازی وقایع در ویدئو روندی است که در آن نوع وقایع یک ویدئو به‌صورت خودکار آشکار می‌شوند. این روند می‌تواند شامل تعیین زمان وقوع هر واقعه در ویدئو و برشمردن وقایع در ویدئو نیز باشد.

در این مقاله، هدف ارائه راه‌کارهایی برای پردازش معنایی ویدئو با توجه به خصوصیات زمانی سیگنال ویدئو است. تشخیص وقایع به‌طورمعمول بسیار پیچیده‌تر از وظایف مشابه مانند تشخیص افراد [1] و تشخیص چهره [2] است. مهم‌ترین دلایل این پیچیدگی عبارتند از: تغییرات بسیار شدید درون‌طبقه‌ای وقایع [3]، طول متغیر

[DOI: 10.52547/jsdp.18.1.134]

ویدئو [4]، نوفه شدید موجود در ویدئو در اثر اعمال فیلترها و پیش‌پردازش‌ها [5]، و نبود ساختار مشخص از پیش تعیین‌شده برای وقایع [6]. درحقیقت، هر واقعه ممکن است در شرایط مختلف شامل پس‌زمینه‌ها، اشخاص، اشیا، و فعالیت‌های مختلف اتفاق بیافتند.

برای غلبه بر این پیچیدگی‌ها، پژوهش‌گران زیادی ویژگی‌های مختلفی مانند ظاهر [7]، [8]، [9]، حرکت [10]، [11]، [12]، [13]، [14] ویژگی‌های صوتی [15]، [16]، [17] را برای بهبود تشخیص وقایع به‌کار برده‌اند.

به‌صورت سنتی ویژگی‌های سطح پایین برای تشخیص وقایع به‌کار می‌رفته‌اند. یکی از دقیق‌ترین ویژگی‌های تصویری برای تشخیص ویدئو IDT [13] است که در آن هم ویژگی‌های مبتنی بر حرکت و هم ویژگی‌های مبتنی بر ظاهر استفاده می‌شود [18]. عملکرد IDT از تمام روش‌های مطرح مبتنی بر ویژگی‌های دست‌ساز [19]، [20] مانند DSIFT [21] و STIP [10] بهتر است [3] و [22]. اما بار محاسباتی سنگین آن، کاربرد این روش را درعمل محدود می‌کند [3]. به‌طوردقیق همانند IDT روش‌های مبتنی بر ویژگی‌های مکانی-زمانی مانند 3DSIFT [23]، MoSIFT [11]، و HoG3D [24] از بار محاسباتی سنگین رنج می‌برند.

با درنظرگرفتن این محدودیت‌ها و همچنین به‌دلیل عملکرد فوق‌العاده شبکه‌های عصبی کانولوشنی[1] (CNNs) [25] در تشخیص تصاویر برای مثال بر روی داده‌های ImageNet [26]، و تصاویر نوفه‌ای [27] پژوهش‌گران درهمین‌اواخر به‌طور گسترده‌ای از CNN در پردازش ویدئو استفاده کرده‌اند [3]، [4]، [28]، [29] و به عملکرد بهتری از لحاظ دقت و نیز بار محاسباتی کمتر دست‌یافته‌اند. یافته‌های جدید نشان می‌دهند که روش‌های مبتنی بر CNN حتی بهترین روش‌های مبتنی بر ویژگی‌های دست‌ساز را پشت سر می‌گذارند [29].

شبکه‌های CNN به شبکه‌های عصبی سنتی شباهت دارند ولی مفاهیم جدیدی مانند لایه کانولوشنی[2] یا لایه واحد یک‌سوشده خطی[3] (ReLu) و نیز روش‌های جدید یادگیری در آن‌ها وجود دارد که عملکرد آن‌ها را در مقایسه با شبکه‌های عصبی سنتی به طرز چشم‌گیری ارتقا می‌دهد [30]. در سال ۲۰۱۲، Krizhevsky و همکاران [30] نخستین CNN با ویژگی‌های جدید را ارائه دادند که توانست برنده مسابقه ILSVRC[4] در سال ۲۰۱۲ [31] شود. این شبکه توانست جهش بزرگی در دقت تشخیص تصاویر به‌وجود آورد و خطای طبقه‌بندی تصاویر در مسابقه یادشده را از ۲۶ درصد به پانزده درصد برساند. بعد از این جهش بود که این شبکه‌ها توجه زیادی را به خود جلب کردند. ساختار CNN شامل چندین نوع لایه است که یک CNN می‌تواند تعداد زیادی از هر کدام از این لایه‌ها داشته باشد. با افزایش تعداد لایه‌ها شبکه عمیق‌تر می‌شود. لایه کانولوشنی [32]، که لایه اصلی شبکه کانولوشنی است، شامل مجموعه‌ای از فیلترها است که کانولوشن آن‌ها با سیگنال ورودی محاسبه و سیگنال حاصل به لایه بعد شبکه داده می‌شود. وزن‌های فیلترها در روند آموزش تعیین می‌شوند. لایه دیگر، لایه ادغام [32] است که نرخ داده را در شبکه کم می‌کند و حجم محاسبات را کاهش می‌دهد. معروف‌ترین نوع لایه ادغام، ادغام بیشینه است که در آن تصویر به قسمت‌های غیرهم‌پوشان تقسیم و مقادیر بیشینه این قسمت‌ها به عنوان خروجی لایه ادغام محاسبه می‌شود.

لایه فعال‌سازی، لایه دیگر موجود در شبکه‌های کانولوشنی است. پراستفاده‌ترین نوع لایه فعال‌ساز، واحد یک‌سوشده خطی [33] است. این لایه مقادیر بزرگتر یا مساوی صفر را بدون تغییر عبور می‌دهد، ولی مقادیر منفی را صفر می‌کند. این لایه موجب می‌شود که روند آموزش شبکه، بدون کاهش محسوس در دقت، بسیار سریع‌تر از زمانی باشد که توابعی همچون سیگموید[5] یا تانژانت هیپربولیک[6] به‌عنوان تابع فعال‌سازی بکار روند [30]. در قسمت‌های انتهایی شبکه یک یا چندلایه تمام‌متصل [30] به‌کار می‌رود. هر نورون در این لایه به تمام نورون‌های لایه قبل متصل است. این لایه در شبکه‌های عصبی سنتی هم وجود دارد. لایه خطا [32] به‌طورمعمول آخرین لایه یک شبکه کانولوشنی است و نحوه تأثیر خطای بین خروجی به‌دست‌آمده شبکه و خروجی مطلوب روی روند آموزش را تعیین می‌کند.

نقشه ویژگی[7] حاصل از لایه‌های مختلف این شبکه‌های از پیش آموزش‌دیده را می‌توان قبل از اعمال به طبقه‌بند نهایی پردازش کرد. پردازش سلسله‌مراتبی معنایی مفاهیم [34] یا پردازش قاب‌ها برای اعمال پردازش چندسطحی به آن‌ها [35] ازجمله این پردازش‌ها هستند. با دنبال‌کردن توصیف سطح قاب، توصیف‌گرها باید کدگذاری شوند. کدگذاری بردار توصیف‌گرهای

[1] Convolutional neural networks
[2] Convolutional layer
[3] Linear rectified unit
[4] ImageNet Large-scale visual recognition competition
[5] Sigmoid
[6] Hyperbolic tangent
[7] Feature map

[DOI: 10.52547/jsdp.18.1.134]

مجتمع محلی یا VLAD[1] [36] ازجمله پرطرفدارترین و بهترین روش‌های کدگذاری است. هرچند در این کدگذاری و کدگذاری‌های مشابه تنها بعد مکانی ویدئو در نظر گرفته‌شده و از بعد زمانی ویدئو در آن‌ها صرف‌نظر می‌شود. به‌دلیل این‌که سیگنال ویدئو هم در بعد زمان و هم در بعد مکان حاوی اطلاعات است، یک کدگذاری خوب می‌تواند علاوه‌بر اطلاعات مکانی، حاوی اطلاعات زمانی ویدئو نیز باشد تا عملکرد تشخیص وقایع بصری را بهبود دهد. با دنبال‌کردن این ایده، در این مقاله یک روش کدگذاری مکانی-زمانی ارائه می‌دهیم که دقت میانگین تشخیص وقایع را به طرز چشم‌گیری افزایش می‌دهد.

نوآوری‌های این مقاله به شرح زیر هستند:

- یک نوآوری فرعی مقاله، تنظیم دقیق یک شبکه CNN از پیش آموزش داده‌شده روی زیرمجموعه‌ای از ویدئوهای مجموعه‌داده CCV است. هدف از این کار بهبود عملکرد تشخیص مفاهیم موجود در قاب‌های ویدئو توسط این شبکه است. با توجه به این‌که شبکه از پیش روی مجموعه‌داده ImageNet آموزش دیده است، ممکن است عملکرد خیلی خوبی روی قاب‌های ویدئو، که از لحاظ کیفیت بصری و نوع مفاهیم تفاوت زیادی با تصاویر ImageNet دارند، نداشته باشد. با تنظیم دقیق، این عیب تا حدی مرتفع می‌شود.
- مهم‌ترین نوآوری‌های این مقاله ارائه یک کدگذاری مکانی- زمانی مبتنی بر VLAD و بیان آن به شکل یک مسأله بهینه‌سازی محدب است؛ سپس، این مسأله با الگوریتم‌های موجود حل مسائل بهینه‌سازی محدب حل می‌شود و نتایج شبیه‌سازی که مؤید کارایی ایده پیشنهادی هستند، ارائه می‌شوند.
- روش پیشنهادی بر روی دو مجموعه‌داده استاندارد و بزرگ ویدئو مورد ارزیابی قرار می‌گیرد و عملکرد آن از لحاظ میانگین دقت متوسط، میانگین فراخوانی متوسط، معیار F و نیز پیچیدگی محاسباتی با بهترین روش‌های موجود مقایسه می‌شود.

روش پیشنهادی ما، مسلماً بار محاسباتی بیشتری نسبت به تشخیص ویدئو مبتنی بر کدگذاری VLAD استاندارد دارد. بر خلاف VLAD استاندارد، کدگذاری پیشنهادی مستلزم حل مسأله بهینه‌سازی و یافتن ضرایب کدگذاری مکانی-زمانی است؛ اما دقت روش پیشنهادی خیلی بیشتر از روش مبتنی بر VLAD استاندارد است. از طرفی بعضی از روش‌های موجود مبتنی بر شبکه‌های حافظه کوتاه‌مدت طولانی[2] (LSTM) [37] دقت بالاتری نسبت به روش مبتنی بر VLAD استاندارد نتیجه می‌دهند، ولی بار محاسباتی بسیار زیادی دارند. هدف ما ارائه روشی است که هم دقت بالاتری را نسبت به VLAD استاندارد نتیجه دهد (دقتی در حد LSTM یا حتی بهتر از آن) و هم بار محاسباتی آن خیلی کمتر از روش‌های مبتنی بر LSTM باشد؛ برای این منظور، شبیه‌سازی‌های زیادی برای مقایسه پیچیدگی محاسباتی روش پیشنهادی با سایر روش‌ها انجام شد که نتایج آن در بخش ۴-۳-۴ قابل مشاهده هستند.

در بخش‌های بعدی این مقاله، در بخش ۲، کارهای اساسی مرتبط با روش پیشنهادی موردبررسی قرار می‌گیرد. بدنه اصلی مقاله را می‌توان در بخش ۳ ملاحظه کرد و نتایج شبیه‌سازی و نتیجه‌گیری نهایی به‌ترتیب در بخش‌های ۴ و ۵ ارائه می‌شوند.

۲- کارهای مرتبط

الگوریتم VLAD [36] نخستین‌بار در سال ۲۰۱۰ به‌عنوان یک نسخه ساده‌شده و غیراحتمالاتی از نمایش کرنل فیشر[3] [38] ارائه شد. این کدگذاری نه‌تنها به دقت بالایی در وظایفی همچون تشخیص تصاویر دست‌یافت، بلکه موجب کاهش قابل‌توجه توان موردنیاز محاسباتی نسبت به روش کرنل فیشر شد. این برتری همچنین در وظایفی همچون جستجوی نمونه[4] [39] و طبقه‌بندی وقایع بصری در ویدئو [40] قابل‌مشاهده است. نسخه‌های مختلفی از کدگذاری اولیه VLAD در ادبیات ارائه شده است که قدرت مدل‌سازی بیشتری نسبت به نسخه اولیه دارند. در مرجع [41]، یک نرمالیزه‌سازی داخلی[5] ارائه شده است تا مشکل مقادیر انفجاری[6] در نمایش VLAD را تا حدی مرتفع کند. برای کاهش مشکل مقادیر انفجاری ویژگی‌های کدشده و برای متعادل ساختن نقش همه توصیف‌گرها در نمایش نهایی تصویر یا ویدئو، نرمالیزه‌کردن باقیمانده[7] نیز ارائه شده است [42].

برای این‌که باقیمانده‌ها به‌طور مساوی در نمایش VLAD نقش داشته باشند، یک روش نرمالیزه‌کردن در [43] ارائه شده است که هر باقیمانده به اندازه نرم ℓ_2

[1] Vector of locally aggregated descriptors

[2] Long short term memory
[3] Fisher
[4] Instance search
[5] Intra normalization
[6] Burstiness
[7] Residual normalization

[DOI: 10.52547/jsdp.18.1.134]

خودش تقسیم می‌شود. این تغییر، باعث بهبود نتیجه نهایی در وظایفی همچون بازیابی تصویر می‌شود. کدگذاری VLAD اصلی، از ادغام میانگین بر روی توصیف‌گرهای کدشده استفاده می‌کند تا توصیف‌گر نهایی حاصل شود. این در حالی است که ادغام بیشینه می‌تواند به‌جای ادغام میانگین استفاده شود تا عملکرد کدگذاری بهبود یابد.

یک روش سرراست برای بهبود عملکرد، استفاده از ترکیبی از هر دو ادغام است [44]. همچنین، روش دیگر استفاده از ترکیب وزن‌دهی‌شده ادغام‌های بیشینه و میانگین است [45]. در مرجع [46] یک ادغام بیشینه‌ای کلی برای VLAD پیشنهاد شده است که اثر حاصل از هم توصیف‌گرهای مکرر و هم توصیف‌گرهای کم‌تکرار را متعادل می‌کند و عملکرد را به‌خصوص زمانی که توصیف‌گرهای کم‌تکرار حاوی اطلاعات ارزشمندی هستند، بهبود می‌دهد.

در [47] کدگذاری تنک[1] و ادغام بیشینه برای کدگذاری ویژگی‌ها و تجمیع آن‌ها در پردازش تصویر استفاده شده‌اند. ادعای نویسندگان این است که روش ارائه‌شده از لحاظ دقت نهایی نسبت به VLAD ترجیح دارد. همچنین نشان داده شده که اضافه‌کردن PCA[2] همراه با سفیدسازی[3] عملکرد VLAD را نسبت به حالت بدون سفیدسازی بهبود داده است. به‌احتمال دلیل این بهبود این است که سفیدسازی اثر اتفاق هم‌زمان توصیف‌گرها را کاهش و درنتیجه عملکرد کلی را بهبود می‌دهد [48].

اگرچه کارهای زیادی در زمینه بهبود VLAD را می‌توان در ادبیات پیدا کرد، بیش‌تر آن‌ها فقط بر روی بعد مکانی ویدئو متمرکز هستند و کار کمی در زمینه بعد زمانی سیگنال ویدئو انجام شده است. مقالاتی هم که به این موضوع پرداخته‌اند، بیشتر به بعد زمانی به‌عنوان اطلاعات اضافی در کنار بعد مکانی برای بهبود کدگذاری متداول VLAD توجه کرده‌اند و هیچ‌یک به یک کدگذاری یا نمایش زمانی مکانی دست نیافته‌اند. برای مدل‌سازی زمانی ویدئو، می‌توان، به‌طور واضح، به‌جای ویژگی‌های مکانی از ویژگی‌های مکانی-زمانی استفاده کرد و سپس VLAD را برای کدگذاری به‌کار گرفت. برای مثال، در [49] نشان داده شده که استفاده از هیستوگرام محلی جریان نوری یا LHOOF[4]، و استفاده از VLAD در تشخیص فعالیت عملکرد بهتری نسبت به استفاده تنها از ویژگی مکانی به همراه کدگذاری VLAD دارد. ترکیب ویژگی‌های مکانی-زمانی با VLAD در تشخیص فعالیت همچنین در [50] موردبررسی قرار گرفته است. اگر قاب‌های یک ویدئو را در فضای VLAD نمایش دهیم، می‌توان ابتدا با بخش‌بندی[5] ویدئو و سپس انتساب یک بردار به هر بخش، به یک نمایش فشرده وابسته به زمان دست‌یافت [51].

برای استفاده از اطلاعات زمانی، همچنین، قاب‌های ویدئو به گروهای قاب یا GoF[6] تقسیم می‌شوند VLAD به‌طور جداگانه به هر GoF اعمال می‌شود [52]. این در حالی است که در به‌کارگیری VLAD استاندارد در کدگذاری ویدئو، یک بردار به کل ویدئو منتسب می‌شود و اطلاعات مکانی دور ریخته می‌شوند. همچنین، توصیف‌گرهای حاصل از VLAD برای هر قاب را می‌توان به حوزه فرکانس برای مثال حوزه فوریه نگاشت کرد تا تغییرات زمانی قاب‌ها در نمایش نهایی لحاظ شوند [53].

درواقع، در هیچ‌کدام از مقالات بالا، بعد زمانی ویدئو به‌طور مستقیم در کدگذاری VLAD داخل نشده است. روش‌های موجود یا از ویژگی‌های زمانی استفاده می‌کنند که در بسیاری موارد، برای مثال در مورد توصیف‌گرهای ثابت سطح قاب مبتنی بر CNN قابل‌اعمال نیستند، یا تبدیل‌ها و بخش‌بندی‌های ساده‌ای را در کنار VLAD بر ویدئوی اصلی اعمال می‌کنند. این روش‌ها بسیار بدیهی هستند و حتی از لحاظ عملکرد نیز بر VLAD اصلی برتری ندارند. برای رفع این عیب، در این مقاله یک کدگذاری VLAD زمان-محور ارائه می‌شود و عملکرد آن موردبررسی قرار می‌گیرد.

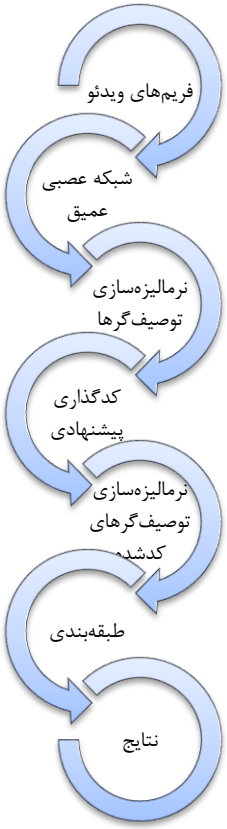
۳- کدگذاری مکانی-زمانی

۳-۱- مدل تشخیص وقایع

فلوچارت سیستم تشخیص وقایع در شکل (۱) نشان داده شده است. ابتدا، یک CNN از پیش آموزش‌دیده به قاب‌هایی که به‌صورت یک‌نواخت از ویدئو نمونه‌برداری شده‌اند اعمال می‌شود. بعد از این‌که مقادیر توصیف‌گر (نمرات مفاهیم) در سطح قاب محاسبه شدند و در لایه آخر شبکه کانولوشنی آماده شدند، نرمالیزه می‌شوند و به بلوک کدگذاری مکانی-زمانی تحویل داده می‌شوند. نرمالیزه‌سازی به‌کاررفته، متداول‌ترین نرمالیزه‌سازی در ادبیات VLAD است و آن هم نرمالیزه‌کردن هر بردار به نرم ℓ_2 آن است.

[1] Sparse coding
[2] Principal component analysis
[3] Whitening
[4] Local Histogram of Oriented Optical Flow
[5] Segmentation
[6] Group of frames

[DOI: 10.52547/jsdp.18.1.134]



(شکل-۱): سیستم تشخیص وقایع بصری: فریم‌های ویدئو با CNN پردازش‌شده و سپس، نرمالیزه و کدگذاری می‌شوند. درنهایت، مجدداً نرمالیزه و طبقه‌بندی می‌شوند.

(Figure-1): Visual event recognition system: Video frames are processed with fine-tuned CNNs, and then are normalized and encoded. Finally, they are normalized again and classified.

۲-۳- کدگذاری VLAD

کدگذاری VLAD به‌عنوان یک نمایش فشرده برای تصویر ارائه شد [36]. همان‌طور که در مقدمه اشاره شد، این نمایش می‌تواند در کدگذاری ویدئو هم مورداستفاده قرار گیرد. در این نمایش، یک کتاب کد $\{c_1, c_2, \dots c_K\}$ متشکل از K واژه تصویری[1] با الگوریتم K-means آموزش داده می‌شود؛ سپس، هر توصیف‌گر در سطح فریم $x_i \in \mathbb{R}^d$ به نزدیک‌ترین کلمه کد منتسب می‌شود. معیار نزدیکی نیز فاصله اقلیدسی در نظر گرفته می‌شود. یعنی خواهیم داشت: $c_j = NN(x_i)$، که در آن NN نزدیک‌ترین همسایه را نشان می‌دهد. درنهایت، در کدگذاری VLAD برای هر یک از واژگان کد، مجموع تفاضل هر توصیف‌گر از نزدیک‌ترین واژه کد یعنی مجموع جملات $x_i - c_j$ ساخته می‌شود؛ بنابراین، توصیف‌گر VLAD به این صورت بیان می‌شود:

$$v_j = \sum_{i:\mathrm{NN}(x_i)=c_j} (x_i - c_j), \quad j = 1, \dots K \qquad (۱)$$

[1] Visual words

سپس، مجموع بردارهای باقیمانده یعنی $v_j \in \mathbb{R}^d$ برای هر واژه کد به هم چسبانده می‌شوند تا یک بردار با اندازه $K \times d$ ساخته شود. درنهایت، بردار حاصل به نرم ℓ_2 نرمالیزه می‌شود تا توصیف‌گر نهایی حاصل شود.

۳-۳- کدگذاری پیشنهادی وابسته به زمان

در این بخش، یک کدگذاری وابسته به زمان VLADپیشنهاد می‌کنیم که نخستین مدل کدگذاری از این نوع است. ابتدا، یک شکل بهینه‌سازی از کدگذاری VLADارائه می‌شود. سپس، تحت یک قید پیوستگی زمانی، مسأله بهینه‌سازی به‌صورت تحلیلی حل می‌شود تا کدگذاری وابسته به زمان موردنظر شکل گیرد.

۱-۳-۳- کدگذاری VLAD به‌عنوان یک مسأله بهینه‌سازی

مشابه الگوریتم اصلیVLAD توصیف‌گرهای سطح قاب را با $X = (x_1, x_2, \dots x_N)$ و مراکز خوشه‌های حاصل از اعمال خوشه‌بندی K-means را با $C = (c_1, c_2, \dots c_K)$ نشان می‌دهیم.

در اینجا، $x_i \in \mathbb{R}^d$ و $c_j \in \mathbb{R}^d$ به ترتیب i امین بردار توصیف‌گر سطح قاب و jامین مرکز خوشه هستند. حال، معادله (۱) را به‌صورت زیر بازنویسی می‌کنیم:

$$v_j = \sum_{i=1}^{N} a_{ij} \qquad (۲)$$

که در آن a_{ij} به‌صورت زیر تعریف می‌شود:

$$a_{ij} = \begin{cases} x_i - c_j & \mathrm{NN}(x_i) = c_j \\ 0 & \text{otherwise} \end{cases} \qquad (۳)$$

چون بردار ستونی $c_j \in \mathbb{R}^d$، j امین مرکز خوشه است، و $C = [c_1, c_2, \dots c_K] \in \mathbb{R}^{d \times K}$ ماتریس حاصل از به هم چسباندن همهٔ مراکز خوشه‌ها است، داریم:

$$a_{ij} = \begin{cases} x_i - C\varphi & \mathrm{NN}(x_i) = c_j \\ 0 & \text{otherwise} \end{cases} \qquad (۴)$$

که در آن φ برداری است که تنها یک درایه یک دارد و بقیه درایه‌های آن صفر هستند. بدیهی است که اگر مؤلفهٔ jام بردار φ برابر یک باشد، یعنی x_i به خوشهj ام تعلق

[DOI: 10.52547/jsdp.18.1.134]

دارد. درواقع، ما دنبال φهایی هستیم که منجر به یافتن نزدیک‌ترین مرکز خوشه به هر یک از توصیف‌گرهای سطح قاب شوند. این موضوع را می‌توان در قالب یک مسأله بهینه‌سازی به‌صورت زیر نوشت:

$$\begin{aligned} &\varphi_i = \min_{\theta} \|x_i - C\varphi\|_2^2 \\ &\text{s.t} \quad \|\varphi\|_0^2 = \|\varphi\|_1^2 = 1 \end{aligned} \tag{۵}$$

که در آن $\|\varphi\|_0$شبه نرم ℓ_0 بردار φ یعنی تعداد درایه‌های ناصفر آن و $\|\varphi\|_1$ نرم ℓ_1 است.

۲-۳-۳- اعمال قید پیوستگی زمانی بر VLAD

تابه‌حال، کدگذاری VLAD را به‌عنوان یک مسأله بهینه‌سازی با مجهول φ_i بیان کردیم که در آن φ_i بیان‌گر اندیس خوشه‌ای است که بردار توصیف‌گر قاب x_i به آن خوشه تعلق یافته است. با درنظرگرفتن VLAD به‌عنوان یک مدل کدگذاری ویدئو با پارامتر φ انتظار داریم که تغییرات موجود در φ در طی قاب‌های متوالی ویدئو کم باشد. تنها در لحظات خاصی ما ممکن است تغییرات کم‌وبیش زیادی در φداشته باشیم. این ایده از اینجا سرچشمه می‌گیرد که یک ویدئوی طبیعی، در شکل غیرفشرده آن، یک سیگنال پیوسته و بسیار همبسته[1] است. درنتیجه، پارامتر مدل یعنی φ در طول یک ویدئو به‌صورت تکه‌ای همواره باقی می‌ماند. شرط بالا را می‌توان به‌عنوان یک جمله جریمه[2] در مسأله بهینه‌سازی VLAD در رابطه (۵) در نظر گرفت. همچنین، چون می‌توان φ را در یک ویدئوی واقعی تکه‌ای ثابت در نظر گرفت، قید زمانی را باید به کل ویدئو اعمال کرد. به‌بیان‌دیگر، یک جمله جریمه ناشی از مجموع تغییرات φ در طول ویدئو را به مسأله بهینه‌سازی می‌افزاییم تا قید پیوستگی زمانی لحاظ شود؛ درنتیجه خواهیم داشت:

$$\begin{aligned} \phi^* = \underset{\substack{\theta_i \\ 1\le i\le N}}{\operatorname{argmin}} & \sum_{i=1}^{N} |x_i - C\varphi_i|_2^2 + \beta \sum_{i=1}^{N-1} |\varphi_{i+1} - \varphi_i|_2^2 \\ \text{s.t.} \quad & |\varphi_i|_0^2 = |\varphi_i|_1^2 = 1, \quad 1 \le i \le N \\ & \lambda \ge 0 \end{aligned} \tag{۶}$$

که در آن λ وزن نسبی جمله جریمه را تعیین می‌کند. $\phi^* \in \mathbb{R}^{K\times N}$ماتریس شامل مقادیر بهینه بردارهای φ_iاست. تابع هزینه در مسأله (۶) تنها شامل نرم فروبنیوس[3] و درنتیجه یک تابع هزینه محدب است؛ ولی وجود نرم ℓ_0 و تساوی برای قید نرم ℓ_1، قیدها را در رابطه (۶) غیرمحدب می‌کند. برای محدب‌سازی قیدها، از رهاسازی[4] نرم ℓ_0 با نرم ℓ_1 و نیز جایگزینی کوچک‌ترمساوی به‌جای تساوی استفاده می‌کنیم. با این کار هر دو قید در رابطه (۶) به یک قید نرم ℓ_1 کوچک‌ترمساوی تبدیل می‌شوند. دوباره با نوشتن مسأله معدل لاگرانژ خواهیم داشت:

$$\begin{aligned} \Theta^* = \underset{\substack{\theta_i \\ 1\le i\le N}}{\operatorname{argmin}} & \sum_{i=1}^{N} |x_i - C\varphi_i|_2^2 + \alpha \sum_{i=1}^{N} |\varphi_i|_1 + \\ & \beta \sum_{i=1}^{N-1} |\varphi_{i+1} - \varphi_i|_2^2 \\ \text{s.t.} \quad & \lambda \ge 0, \gamma \ge 0 \end{aligned} \tag{۷}$$

رابطه (۲) یک مسأله بهینه‌سازی محدب در شکل دوگان لاگرانژ است که حل آن با ابزارهای بهینه‌سازی محدب ممکن است. ما برای حل آن از کتابخانه CVX [54] در نرم‌افزار MatLab استفاده کردیم. CVX یک کتابخانه برای حل مسائل بهینه‌سازی محدب است. با یافتن φ_iها با حل معادله (۲) از طریق کتابخانه CVX، مقادیر بهینه φ_iها را در روابط (۲ و ۴) جایگذاری می‌کنیم تا نمایش VLAD مکانی-زمانی به‌دست آید.

۴- نتایج تجربی

۴-۱- پیاده‌سازی

ما در این مقاله، از دو مجموعه‌دادهٔ CCV[5] و ActivityNet برای آزمودن روش پیشنهادی استفاده می‌کنیم. مجموعه‌دادهٔ CCV [55] از بیست ردهٔ معنایی شامل ۹۳۱۷ ویدئو از یوتیوب تشکیل شده است. این مجموعه‌داده به دو بخش آموزش و آزمون شامل ۴۶۵۹ ویدئوی آموزشی و ۴۶۵۸ ویدئوی آزمون تقسیم شده است.

مجموعه‌دادهٔ ActivityNet شامل ۶۴۸ ساعت ویدئوی یوتیوب است. این مجموعه شامل حدود بیست‌هزار ویدئو و دویست رده است. به‌طور متوسط در هر ویدئو ۱/۵۴ فعالیت مختلف وجود دارد، به‌عبارت‌دیگر، هر ویدئو ممکن است شامل بیش از یک موضوع باشد. در بخش‌بندی اصلی، زیربخش‌های آموزش، آزمون، و ارزیابی به ترتیب شامل ۵۰، ۲۵، و ۲۵ درصد از کل مجموعه‌داده هستند. در این مقاله از نسخه ۱/۳[6] این مجموعه‌داده استفاده شده است.

[1] Correlated
[2] Penalty term
[3] Frobenius
[4] Relaxation
[5] Columbia consumer video
[6] Version 1.3

[DOI: 10.52547/jsdp.18.1.134]

همچنین، برای استخراج مفاهیم از یک شبکه با ساختار ارائه‌شده در [56]، شامل پنج لایه کانولوشنی و سه لایه تمام متصل[1] استفاده شده است. این مدل از قبل روی مجموعه‌داده ImageNet آموزش داده شده است. این شبکه سپس بر روی زیرمجموعه‌ای از قاب‌های ویدئوی مجموعه داده CCV تنظیم شده است. برای طبقه‌بندی مبتنی بر SVM کتابخانه LIBSVM [57] استفاده می‌شود. کتابخانه VideoUtils [58] برای کدگشایی و خواندن فایل‌های ویدئویی به‌کار می‌رود. کتابخانه MatConvNet [59] نیز برای اعمال CNN بر قاب‌های ویدئو مورد استفاده قرار می‌گیرد. VLFeat [60] برای انتساب خوشه‌ها، کدگذاری VLAD استاندارد (روش مبنا)، و نرمالیزه‌کردن توصیف‌گرهای کدشده به‌کار می‌رود. ویدئوها با نرخ پایین یک قاب بر ثانیه نمونه‌برداری می‌شوند تا محاسبات سریع‌تر شود. مطابق مرجع [61]، استفاده از تنها یک قسمت کوچک از قاب‌ها، دقت میانگین تشخیص ویدئو را کاهش نمی‌دهد.

۴-۲- معیارهای ارزیابی

در این مقاله، برای ارزیابی روش پیشنهادی و مقایسه آن با بهترین روش‌های موجود از سه معیار ارزیابی استفاده می‌شود. این سه معیار به‌طور گسترده‌ای برای ارزیابی روش‌های طبقه‌بندی مورد استفاده قرار می‌گیرند.

۴-۲-۱- معیار میانگین دقت متوسط[2] (mAP)

دقت متوسط[3] تشخیص وقایع برای هر رده از وقایع به‌صورت نسبت تعداد وقایع بازیابی‌شده[4] مرتبط (وقایعی که تعلق آن‌ها به رده مورد نظر به درستی پیش‌بینی شده است) به کل وقایع بازیابی‌شده برای آن رده تعریف می‌شود [62]. میانگین دقت متوسط (mAP) نیز با میانگین‌گیری روی مقادیر دقت متوسط تمام رده‌ها به‌دست می‌آید.

۴-۲-۲- معیار میانگین فراخوانی متوسط[5] (mAR)

فراخوانی متوسط[6] تشخیص وقایع برای هر رده از وقایع به‌صورت نسبت تعداد وقایع بازیابی‌شده مرتبط (وقایعی که تعلق آن‌ها به رده مورد نظر به درستی پیش‌بینی شده است) به کل وقایع مرتبط (مجموع وقایع بازیابی‌شده مرتبط و وقایع بازیابی‌نشده مرتبط) به آن رده تعریف می‌شود [62]. میانگین فراخوانی متوسط (mAR) نیز با میانگین‌گیری روی مقادیر فراخوانی متوسط تمام رده‌ها به‌دست می‌آید.

۴-۲-۳- معیار F[7]

معیار F به صورت میانگین هارمونیکی بین mAP و mAR تعریف می‌شود [63]:

$$F = 2\frac{mAP.mAR}{mAP + mAR} \qquad (۸)$$

اگر هر کدام از مقادیر mAP یا mARکوچک باشند، مقدار F نیز کوچک می‌شود. مقدار F وقتی زیاد است که mAP و mAR به طور هم‌زمان مقدار بزرگی داشته باشند.

۴-۳- نتایج شبیه‌سازی

۴-۳-۱- انتخاب پارامتر

برای انتخاب پارامترهای بهینه شامل تعداد مؤلفه‌های اصلی، تعداد خوشه‌ها، اندازه واژگان[8]، و مقدار پارامترهای منظم‌سازی[9] α و β، اعتبارسنجی متقابل روی داده‌های اعتبارسنجی مجموعه‌داده CCV انجام شده است. بنابراین، مقادیر مختلف پارامترها تغییر می‌یابند تا مقادیر بهینه با توجه به دقت طبقه‌بندی وقایع روی داده‌های اعتبارسنجی به‌دست آیند.

۴-۳-۱-۱- بعد توصیف‌گرها

برای ایجاد تعادلی بین سرعت و دقت الگوریتم، می‌توان از PCA همراه با سفیدسازی برای کاهش ابعاد توصیف‌گرها استفاده کرد. جدول (۱) اثر کاهش بعد توصیف‌گر بعد از اعمال PCA بر روی دقت الگوریتم تشخیص وقایع مبتنی بر کدگذاری VLAD بر روی تعداد ثابتی خوشه برای مجموعه‌داده CCV را نشان می‌دهد. همان‌طور که دیده می‌شود، مقادیر دقت میانگین با کاهش بعد توصیف‌گر به‌صورت یک‌نواخت کاهش زیادی می‌یابند؛ بنابراین، ما از همان توصیف‌گرهای 1000 بعدی با وجود کاهش سرعت شبیه‌سازی‌ها استفاده می‌کنیم.

(جدول-۱): دقت میانگین یا mAP[10] روش مبنای VLAD بعد از کاهش بعد روی مجموعه‌داده CCV با تعداد واژگان ۱۶.

(Table-1): Mean average precision of standard VLAD after dimensionality reduction on CCV with vocabulary size of 16.

بعد	32	64	128	256	512	1000
mAP(%)	64.5	65.3	66.7	67.9	68.6	70.6

[1] Fully connected
[2] Mean average precision
[3] Average precision
[4] Retrieved events
[5] Mean average recall
[6] Average recall
[7] F-measure
[8] Vocabulary size
[9] Regularization
[10] Mean average Precision

[DOI: 10.52547/jsdp.18.1.134]

(جدول-۲): اثر تعداد خوشه‌ها بر دقت میانگین VLAD با توصیف‌گرهای هزار بُعدی روی مجموعه‌داده CCV.

(Table-2): Effect of number of clusters on mAP of VLAD with 1000 dimensional descriptors in CCV dataset.

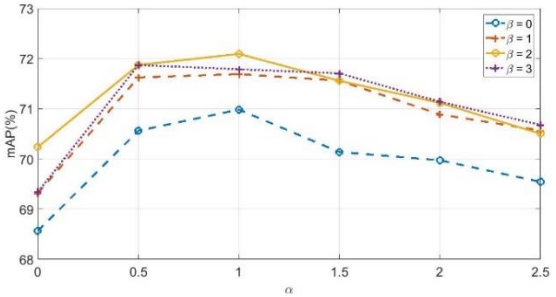
K	1	2	4	8	16	32	64	128
mAP(%)	66.8	70.6	72.0	72.4	70.6	68.1	67.0	66.2

۲-۱-۳-۴- تعداد خوشه‌ها

جدول (۲) اثر تعداد خوشه‌ها در الگوریتم مبنای VLAD بر روی دقت میانگین تشخیص وقایع روی مجموعه‌داده CCV را نشان می‌دهد. همان‌طور که از جدول دیده می‌شود، افزایش K دقت میانگین را افزایش می‌دهد، اما افزایش بیشتر K به بیشتر از هشت، دقت تشخیص را بهتر نمی‌کند. درواقع، وقتی تعداد خوشه‌ها افزایش زیادی پیدا می‌کند، توصیف‌گرهای کدشده خیلی تنک می‌شوند. طبقه‌بندهای معمول مانند SVM عملکرد چندان خوبی روی توصیف‌گرهای تنک ندارند [34] و [64]، که این امر موجب کاهش دقت تشخیص می‌شود. به همین دلیل است که روش‌های نرمالیزه‌سازی زیادی مانند آنچه در [34] و [64] آمده است، برای کاهش اثر تنکی پیشنهاد شده‌اند.

۳-۱-۳-۴- انتخاب پارامترهای منظم‌سازی

همان‌طور که در رابطه نهایی کدگذاری مکانی-زمانی پیشنهادی دیده می‌شود، این کدگذاری تابعی از پارامترهای منظم‌سازی، یعنی α و β است. شکل (۲) اثر این پارامترها بر دقت میانگین تشخیص وقایع را روی مجموعه‌داده CCV برای توصیف‌گرهای هزار بعدی با هشت مرکز خوشه نشان می‌دهد. همان‌طور که از این شکل برمی‌آید، مقادیر $\alpha=1$ و $\beta=2$ منجر به بیشترین مقدار دقت میانگین می‌شوند. این مقادیر، وزن‌های بهینه جملات جریمه ناشی از تنکی توصیف‌گرها و تنکی در حوزه تفاضل زمانی توصیف‌گرها را نشان می‌دهند. با توجه به مقادیر به‌دست‌آمده، وزن (اهمیت) تنکی توصیف‌گرها در حوزه تفاضل دو برابر وزن (اهمیت) تنکی خود توصیف‌گرها است. این مسأله با شهود ما مبنی بر اهمیت بسیار زیاد لحاظ‌کردن ارتباط زمانی قاب‌ها مطابقت دارد. کمینه مقدار دقت میانگین در شکل (۲) هنگامی است که $\alpha=\beta=0$ و این مساوی دقت میانگین در کدگذاری استاندارد VLAD است. این موضوع به این دلیل است که وقتی ضریب متناظر پیوستگی زمانی را صفر قرار می‌دهیم، کدگذاری پیشنهادی در رابطه (۷) به کدگذاری استاندارد VLAD کاهش می‌یابد؛ بنابراین برای $\alpha=\beta=0$، دقت روش پیشنهادی برابر روش VLAD مبنا می‌شود.

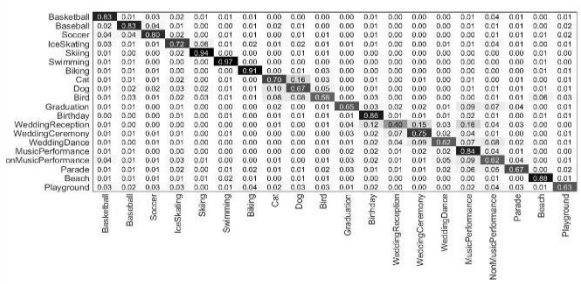


(شکل-۲): دقت تشخیص وقایع نسبت به پارامترهای تنظیم برای توصیف‌گرهای هزار بُعدی و هشت خوشه روی مجموعه‌داده CCV.

(Figure-2): Mean average precision with respect to regularization prameters for 1000 dimensional descriptors and eight clusters on CCV dataset.

۲-۳-۴- مقایسه نتایج بر روی مجموعه‌داده CCV

شکل (۳) ماتریس ابهام طبقه‌بندی[1] برای مجموعه‌داده CCV را نشان می‌دهد. ردیف‌های این ماتریس متناظر با رده‌های واقعی هستند درحالی‌که ستون‌های آن متناظر با رده‌های پیش‌بینی‌شده توسط الگوریتم تشخیص وقایع هستند. مقادیر ماتریس نرمالیزه شده‌اند تا به‌جای دفعات رخداد بیان‌گر احتمال باشند. درایه ردیف i ام و ستون j ام احتمال این را نشان می‌دهد که رده واقعه iام رده واقع jام طبقه‌بندی شود. مقادیر خارج از قطر اصلی ماتریس ابهام، خطای پیش‌بینی را نشان می‌دهند. همان‌طور که انتظار می‌رود، وقایعی که همبستگی مفهومی و بصری زیادی باهم دارند، مشکل‌تر از هم جدا می‌شوند. با توجه به شکل (۳) می‌توان دید که وقایع مرتبط با عروسی یعنی Wedding reception، Wedding Ceremony که شباهت زیادی از لحاظ بصری و مفهومی به هم دارند، در بعضی از ویدئوها با هم اشتباه گرفته می‌شوند. این موضوع موجب ایجاد مقدار بزرگ خارج از قطر اصلی در این ناحیه می‌شود. به‌طوردقیق همین موضوع در مورد دو رده Cat و Dog نیز صحیح است که به‌دلیل شباهت مفهومی به هم، در بعضی از ویدئوها اشتباه تشخیص داده می‌شوند.



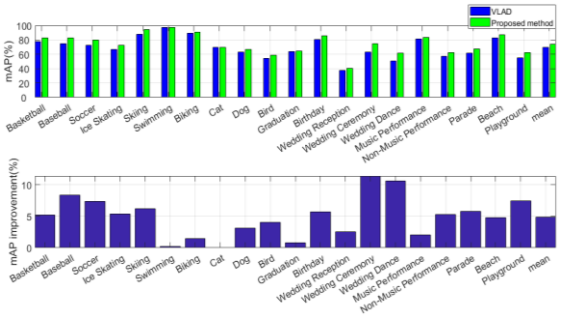
(شکل-۳): ماتریس ابهام طبقه‌بندی روش پیشنهادی بر روی مجموعه داده CCV.

(Figure-3): Classification confusion matrix for proposed method on CCV dataset.

[1] Classification confusion matrix

[DOI: 10.52547/jsdp.18.1.134]

شکل (۴) دقت طبقه‌بندی برای هر رده را به‌صورت مجزا نشان می‌دهد. همان‌طور که دیده می‌شود، کدگذاری مکانی-زمانی ارائه‌شده دقت را بر روی تمام رده‌های وقایع بهبود می‌دهد. میزان بهبود دقت متوسط در هر رده تابعی از میزان تأثیر درنظرگرفتن ارتباط زمانی قاب‌ها در توصیف ویدئو است. هرچقدر تأثیر اطلاعات زمانی در توصیف یک رده از وقایع بیشتر باشد، میزان بهبود نسبی روش پیشنهادی بیشتر خواهد بود. برای مثال همان‌طور که در شکل (۴) دیده می‌شود، تأثیر اطلاعات زمانی در رده Cat بسیار ناچیز بوده است ولی رده Wedding ceremony بهره بسیار زیادی از اطلاعات زمانی (ارتباط زمانی قاب‌ها) برده است. به‌طور متوسط روی کل مجموعه‌داده CCV، ۴/۹ درصد بهبود در میانگین دقت متوسط داشته‌ایم.



(شکل-۴): دقت متوسط طبقه‌بندی حاصل از کدگذاری پیشنهادی برای هر کلاس روی مجموعه داده CCV.

(Figure-4): Per-class classification average precision of the proposed method on CCV dataset.

۴-۳-۳- مقایسه نتایج بر روی مجموعه‌داده ActivityNet

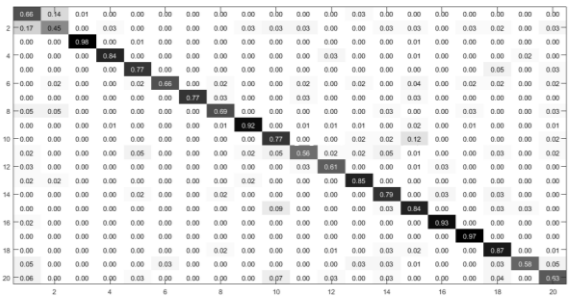
شکل (۵) ماتریس ابهام طبقه‌بندی برای مجموعه‌داده ActivityNet را نشان می‌دهد. به‌دلیل تعداد زیاد رده‌های این مجموعه‌داده، ماتریس ابهام طبقه‌بندی تنها برای بیست رده از این مجموعه‌داده که به‌صورت تصادفی انتخاب شده‌اند، رسم شده است تا قابل رؤیت باشد. ردیف‌های این ماتریس متناظر با رده‌های واقعی هستند درحالی‌که ستون‌های آن متناظر با رده‌های پیش‌بینی‌شده توسط الگوریتم تشخیص وقایع هستند. مقادیر ماتریس نرمالیزه شده‌اند تا به‌جای دفعات رخداد بیان‌گر احتمال باشند. درایه ردیف i ام و ستون j ام احتمال این را نشان می‌دهد که رده واقعه i ام رده واقع jام طبقه‌بندی شود. مقادیر خارج از قطر اصلی ماتریس ابهام، خطای پیش‌بینی را نشان می‌دهند. همانند مجموعه‌داده CCV، در اینجا نیز وقایعی که همبستگی مفهومی و بصری زیادی باهم دارند، مشکل‌تر از هم جدا می‌شوند. با توجه به شکل (۵) می‌توان دید که وقایع E1 و E2 شباهت زیادی از لحاظ بصری و مفهومی به هم دارند و در بعضی از ویدئوها با هم اشتباه گرفته می‌شوند. این موضوع موجب ایجاد مقدار بزرگ خارج از قطر اصلی در این ناحیه می‌شود. به‌طوردقیق همین موضوع در مورد دو رده E10 و E15 نیز صحیح است که به‌دلیل شباهت مفهومی به هم، در بعضی از ویدئوها اشتباه تشخیص داده می‌شوند.

جدول (۳) عملکرد کدگذاری مکانی-زمانی ارائه‌شده را با بهترین روش‌های موجود مقایسه می‌کند. همان‌طور که نشان داده شده است، کدگذاری وابسته به زمان ارائه‌شده بهترین عملکرد را دارد. این در حالی است که بیش‌تر روش‌های دیگر از ویژگی‌های صوتی نیز در کنار ویژگی‌های تصویری استفاده می‌کنند؛ اما ورودی کدگذاری ارائه‌شده تنها ویژگی‌های تصویری است. این افزایش دقت ناشی از لحاظ‌کردن ساختاری اطلاعات زمانی (ارتباط زمانی قاب‌ها) در روند کدگذاری توصیف‌گرها است. همان‌طورکه در جدول (۳) دیده می‌شود، روش پیشنهادی روی مجموعه‌داده CCV، میانگین دقت متوسط، میانگین فراخوانی متوسط، و معیار F را به‌ترتیب ۲/۹، ۴/۴ و ۳/۷ درصد بهبود می‌دهد.

(جدول-۳): مقایسه نتایج روش پیشنهادی با بهترین روش‌های موجود برحسب میانگین دقت متوسط، میانگین فراخوانی متوسط، و معیار F بر روی مجموعه‌داده CCV.

(Table-3): Comparison of the results between proposed method and the state of the art in terms of mAP, mAR, and F-measure on CCV dataset.

روش	mAP (%)	mAR (%)	F (%)
Jiang و همکاران [55]	59.5	57.3	58.3
Xu و همکاران [65]	60.3	57.4	58.8
Ma و همکاران [66]	63.0	58.1	60.4
Yeو همکاران [67]	64.0	62.3	63.1
Jhuo و همکاران [68]	64.0	62.1	63.0
Liu و همکاران [69]	68.2	66.5	67.3
Zhao و همکاران [37]	69.1	65.9	67.4
Wuو همکاران [70]	70.6	67.8	69.1
روش پیشنهادی	**73.5**	**72.2**	**72.8**

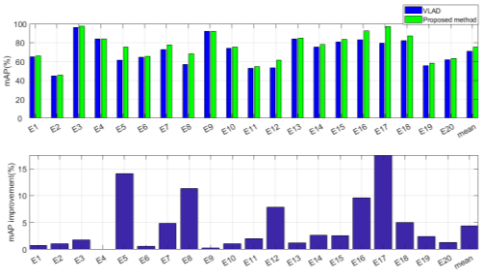


(شکل-۵): ماتریس ابهام طبقه‌بندی روش پیشنهادی بر روی مجموعه داده ActivityNet.

(Figure-5): Classification confusion matrix for proposed method on ActivityNet dataset.

[DOI: 10.52547/jsdp.18.1.134]

شکل (۶) دقت طبقه‌بندی برای هر کلاس را به‌صورت مجزا نشان می‌دهد. همان‌طور که دیده می‌شود، کدگذاری مکانی-زمانی ارائه‌شده دقت را بر روی تمام رده‌های وقایع بهبود می‌دهد. میزان بهبود دقت متوسط در هر رده در اینجا نیز مانند مجموعه‌داده CCV تابعی از میزان تأثیر درنظرگرفتن ارتباط زمانی قاب‌ها در توصیف ویدئو است. هرچقدر تأثیر اطلاعات زمانی در توصیف یک رده از وقایع بیشتر باشد، میزان بهبود نسبی روش پیشنهادی بیشتر خواهد بود. برای مثال همان‌طورکه در شکل (۶) دیده می‌شود، تأثیر اطلاعات زمانی در رده‌های E4 و E9 بسیار ناچیز بوده است، ولی رده E17 بهره بسیار زیادی از اطلاعات زمانی (ارتباط زمانی قاب‌ها) برده است. به‌طور متوسط روی کل مجموعه‌داده ActivityNet، ۴/۴ درصد بهبود در میانگین دقت متوسط داشته‌ایم.



(شکل-۶): دقت متوسط طبقه‌بندی حاصل از کدگذاری پیشنهادی برای هر کلاس روی مجموعه داده ActivityNet.

(Figure-6): Per-class classification average precision of the proposed method on ActivityNet dataset.

(جدول-۴): مقایسه نتایج روش پیشنهادی با بهترین روش‌های موجود برحسب میانگین دقت متوسط، میانگین فراخوانی متوسط، و معیار F بر روی مجموعه‌داده ActivityNet.

(Table-4): Comparison of the results between proposed method and the state of the art in terms of mAP, mAR, and F-measur on ActivityNet dataset.

روش	mAP (%)	mAR (%)	F (%)
Jiang و همکاران [55]	63.8	62.7	63.2
Xu و همکاران [65]	64.4	62.1	63.2
Ma و همکاران [66]	66.5	63.3	64.8
Yeو همکاران [67]	67.8	66.5	67.1
Jhuo و همکاران [68]	68.3	68.0	68.1
Liu و همکاران [69]	71.5	69.3	70.3
Zhao و همکاران [37]	72.1	70.0	71.0
Wuو همکاران [70]	73.0	71.4	72.1
روش پیشنهادی	**75.4**	**73.6**	**74.4**

جدول (۴) عملکرد کدگذاری مکانی-زمانی ارائه‌شده را با بهترین روش‌های موجود مقایسه می‌کند. همان‌طورکه نشان داده شده است، کدگذاری وابسته به زمان ارائه‌شده بهترین عملکرد را دارد. این در حالی است که بیش‌تر روش‌های دیگر از ویژگی‌های صوتی نیز در کنار ویژگی‌های تصویری استفاده می‌کنند؛ اما ورودی کدگذاری ارائه‌شده تنها ویژگی‌های تصویری است. این افزایش دقت ناشی از لحاظ‌کردن ساختاری ارتباط زمانی قاب‌ها در روند کدگذاری توصیف‌گرها است. همان‌طورکه در جدول (۴) دیده می‌شود، روش پیشنهادی روی مجموعه‌داده ActivityNet، میانگین دقت متوسط، میانگین فراخوانی متوسط، و معیار F را به‌ترتیب ۲/۴، ۲/۲ و ۲/۳ درصد بهبود می‌دهد.

۴-۳-۴- بازدهی زمان اجرا

مطابق مرجع [37]، از بازدهی زمان اجرا برای مقایسه پیچیدگی محاسباتی روش پیشنهادی با بهترین روش‌های موجود استفاده می‌کنیم. برای این کار زمان مورد نیاز برای پردازش یک ثانیه از ویدئوهای آزمون را برای روش پیشنهادی با بهترین روش‌های موجود مقایسه می‌کنیم. ویدئوها بر روی یک رایانه با پردازنده Intel® Core™ i5-2430M و پردازنده گرافیکی NVIDIA GTX1050 پردازش شده‌اند. سه روش با کمترین پیچیدگی محاسباتی (بیشترین طول پردازش‌شده ویدئو در یک ثانیه) با فونت ضخیم‌تر نشان داده شده‌اند.

همان‌طورکه در جدول (۵) دیده می‌شود، روش پیشنهادی پس از روش VLAD مبنا کمترین پیچیدگی محاسباتی را دارد. روش پیشنهادی حدود ۲/۱۶ برابر سریع‌تر از بهترین روش موجود است. این در حالی است که روش پیشنهادی دقت بالاتری را نیز در تشخیص وقایع دارد. بنابراین؛ روش ارائه‌شده هم از لحاظ دقت و هم از لحاظ پیچیدگی محاسباتی بهترین روش‌های موجود را پشت سر می‌گذارد.

(جدول-۵): مقایسه روش پیشنهادی با بهترین روش‌های موجود برحسب میانگین طول قابل پردازش ویدئو (بر حسب ثانیه) در یک ثانیه.

(Table-5): Comparison of the proposed method and the state of the art in terms of average length of videos (in seconds) which can be processed per second.

روش	طول متوسط ویدئو (بر حسب ثانیه) که در یک ثانیه قابل پردازش است
روش VLAD مبنا	**56.3**
Xu و همکاران [65]	12.1
Ma و همکاران [66]	11.5
Yeو همکاران [67]	14.2

[DOI: 10.52547/jsdp.18.1.134]

Jhuo و همکاران [68]	13.6
Liu و همکاران [69]	17.8
Zhao و همکاران [37]	19.5
Wuو همکاران [70]	**22.3**
روش پیشنهادی	**48.1**

۵- نتیجه‌گیری

در این مقاله، به بررسی تحلیل وقایع بصری در ویدئو پرداختیم. توصیف‌گرهای مبتنی بر یادگیری عمیق را برای توصیف اولیه قاب‌ها به‌کار بردیم. سپس، یک روش کدگذاری مکانی-زمانی برای توصیف ویدئو ارائه دادیم. این روش، که مبتنی بر کدگذاری VLAD است، بعد زمانی سیگنال ویدئو را در مدل آن لحاظ می‌کند و به طبقه‌بندی بهتری برای تشخیص وقایع بصری منجر می‌شود. روش ارائه‌شده را در قالب یک مسأله بهینه‌سازی محدب بیان و با کتابخانه‌های موجود مسائل محدب آن را حل کردیم. درنهایت، روش ارائه‌شده را روی دو مجموعه‌داده عمومی و بزرگ آزمودیم. روش پیشنهادی نه‌تنها عملکرد تشخیص وقایع بصری را با سه معیار میانگین دقت متوسط، میانگین فراخوانی متوسط، و معیار F نسبت به بهترین روش‌های موجود بهبود می‌دهد، بار محاسباتی کمتری از آنها نیز دارد. در ادامه این پژوهش مطلوب است که کدگذاری مکانی-زمانی پیشنهادی را به‌صورت یک لایه شبکه CNN تبدیل کرد. با این کار می‌توان کل روند تشخیص مفاهیم و کدگذاری آنها و یافتن توصیف‌گر ویدئو را با اعمال یک CNN به ویدئو به انجام رساند. با انجام این کار، استخراج ویژگی‌های سطح قاب و کدگذاری آن‌ها در روندی توأم با هم به‌دست می‌آیند. پیش‌بینی می‌شود که این روند نسبت به روند فعلی که این دو کار مستقل از هم صورت می‌گیرند به نتایج بهتری هم از لحاظ دقت و هم از لحاظ بار محاسباتی منجر شود.

۶- مراجع 6- References

[1] M. Mohseni and M. Seriani, "Pedestrian Detection in Infrared Image Sequences Using SVM and Histogram Classifiers," *JSDP*, vol. 6, no. 1, pp. 79–90, 2009.

[۱] محسن، سریانی محسن. تشخیص عابر پیاده با استفاده از کلاس بندهای SVM و هیستوگرام در توالی تصاویر مادون قرمز. پردازش علائم و داده‌ها. ۱۳۸۸; ۶ (۱).

[2] S. Shafeipour Yourdeshahi, H. Seyedarabi, and A. Aghagolzadeh, "Video based Face Recognition Using Orthogonal Locality Preserving Projection," *JSDP*, vol. 13, no. 2, pp. 139–149, Sep. 2016.

[۲] شفیع‌پور یوردشاهی سجاد، سیدعربی هادی، آقاگل‌زاده علی. شناسایی چهره در رشته‌های ویدیویی با استفاده از افکنش متعامد با حفظ ساختار محلی . پردازش علائم و داده‌ها. ۱۳۹۵; ۱۳ (۲): ۱۳۹-۱۴۹.

[3] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015*, pp. 1798–1807, Accessed: Apr. 08, 2016. [Online].

[4] L. Wang, C. Gao, J. Liu, and D. Meng, "A novel learning-based frame pooling method for event detection," *Signal Processing*, vol. 140, pp. 45–52, 2017.

[5] S. Kwak, B. Han, and J. H. Han, "Scenario-based video event recognition by constraint flow," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3345–3352.

[6] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2005, vol. 1, pp. 886–893.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] K. E. Van De Sande, T. Gevers, and C. G. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[10] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[11] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," Carnegie Mellon University, Technical Report CMU-CS-09-161, 2009.

[12] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 3169–3176.

[DOI: 10.52547/jsdp.18.1.134]

[13] H. Wang and C. Schmid, "Action recognition with improved trajectories," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 3551–3558.

[14] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with Fisher vectors on a compact feature set," *in Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2013, pp. 1817–1824.

[15] F. Metze, S. Rawat, and Y. Wang, "Improved audio features for large-scale multimedia event detection," *in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2014, pp. 1–6.

[16] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, 2008.

[17] V. S. Tseng, J.-H. Su, J.-H. Huang, and C.-J. Chen, "Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 260–267, 2008.

[18] X. Peng and C. Schmid, "Encoding feature maps of cnns for action recognition," presented at the CVPR, THUMOS Challenge Workshop, 2015, Accessed: Apr. 08, 2016. [Online].

[19] R. Baradaran, E.Golpar-Raboki, "Feature Extraction and Efficiency Comparison Using Dimension Reduction Methods in Sentiment Analysis Context", *JSDP*, 2019, vol. 16 (3), pp. 88-79.

[۱۹] برادران راضیه، گلپر رابوکی عفت. استخراج ویژگی و بررسی کارآیی روش‌های کاهش بُعد در زمینه تحلیل احساس. پردازش علائم و داده‌ها. ۱۳۹۸; ۱۶ (۳): ۷۹-۸۸

[20] F. Sherafati, J.Tahmoresnezhad, "Image Classification via Sparse Representation and Subspace Alignment", *JSDP*, 2020, vol.17 (2) , pp.58-47

[۲۰] شرافتی فریماه، طهمورث نژاد جعفر. طبقه‌بندی تصاویر با استفاده از نمایش تُنُک و تطبیق زیرفضا. پردازش علائم و داده‌ها. ۱۳۹۹; ۱۷ (۲): ۵۸-۴۷.

[21] R. Aly et al., "The AXES submissions at TrecVid 2013," 2013.

[22] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[23] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *in Proceedings of the 15th ACM international conference on Multimedia*, Sep. 2007, pp. 357–360.

[24] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," i*n Proceedings of the 19th British Machine Vision Conference (BMVC)*, 2008, pp. 99.1-99.10.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2009, pp. 248–255.

[27] M. Momeny, M .A Sarram, A. Latif, R. Sheikhpour, "A Convolutional Neural Network based on Adaptive Pooling for Classification of Noisy Images", *JSDP*, 2021, vol.17 (4), pp.139-154.

[۲۷] مومنی محمد، صرام مهدی آقا، لطیف علی محمد، شیخ‌پور راضیه. ارائه یک شبکه عصبی کانولوشنال مبتنی بر ادغام تطبیقی پویا برای طبقه‌بندی تصاویر نوفه‌ای. پردازش علائم و داده‌ها. ۱۳۹۹; ۱۷ (۴): ۱۳۹-۱۵۴

[28] D. Oneata, J. Verbeek, and C. Schmid, "The LEAR submission at Thumos 2014," 2014.

[29] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting image-trained cnn architectures for unconstrained video classification," i*n Proceedings of the 26th British Machine Vision Conference*, 2015, p. 60.1-60.13.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Proceedings of the Advances in neural information processing systems*, Jan. 2012, pp. 1097–1105.

[31] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, "ILSVRC-2012, 2012," 2012, [Online]. Available: http://www. image-net. org/challenges/LSVRC.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *in Proceedings of the European Conference on Computer Vision*, 2014, pp. 818–833.

[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *in Proceedings of the 27th International*

[DOI: 10.52547/jsdp.18.1.134]

[45] T. De Campos, G. Csurka, and F. Perronnin, “Images as sets of locally weighted features,” *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 68–85, 2012.

[46] N. Murray and F. Perronnin, “Generalized Max Pooling,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 2473–2480, doi: 10.1109/CVPR.2014.317.

[47] T. Ge, Q. Ke, and J. Sun, “Sparse-Coded Features for Image Retrieval.,” *in Proceedings of the British Machine Vision Conference (BMVC)*, 2013, pp. 1–11, Accessed: Nov. 10, 2016. [Online].

[48] H. Jégou and O. Chum, “Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening,” *in Proceedings of the European Conference on Computer Vision (ECCV)*, 2012, pp. 774–787, Accessed: Nov. 10, 2016. [Online].

[49] M. K. Reddy, S. Arora, and R. V. Babu, “Spatio-temporal feature based VLAD for efficient video retrieval,” *in Proceedings of 4th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2013, pp. 1–4, Accessed: Nov. 10, 2016. [Online].

[50] M. Jain, H. Jegou, and P. Bouthemy, “Better Exploiting Motion for Better Action Recognition,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 2555–2562, doi: 10.1109/CVPR.2013.330.

[51] M. Douze, H. Jégou, C. Schmid, and P. Pérez, “Compact video description for copy detection with precise temporal alignment,” *in Proceedings of the European Conference on Computer Vision, 2010*, pp. 522–535, Accessed: Nov. 10, 2016. [Online].

[52] A. Abbas, N. Deligiannis, and Y. Andreopoulos, “Vectors of locally aggregated centers for compact video representation,” *in 2015 IEEE International Conference on Multimedia and Expo (ICME)*, 2015, pp. 1–6, Accessed: Nov. 07, 2016. [Online].

[53] J. Revaud, M. Douze, C. Schmid, and H. Jegou, “Event Retrieval in Large Video Collections with Circulant Temporal Encoding,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 2459–2466, doi: 10.1109/CVPR.2013.318.

[54] M. Grant and S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, 2014.

[55] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, “Consumer video understanding: A benchmark database and an evaluation of human and machine performance,” *in Proceedings of the 1st ACM International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[34] M. Soltanian and S. Ghaemmaghami, “Hierarchical Concept Score Post-processing and Concept-wise Normalization in CNN based Video Event Recognition,” *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 157–172, 2019.

[35] Y. Han, X. Wei, X. Cao, Y. Yang, and X. Zhou, “Augmenting image descriptions using structured prediction output,” *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1665–1676, 2014.

[36] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3304–3311.

[37] Z. Zhao, Y. Song, and F. Su, “Specific video identification via joint learning of latent semantic concept, scene and temporal structure,” *Neurocomputing*, vol. 208, pp. 378–386, 2016.

[38] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8, Accessed: Jul. 06, 2016. [Online].

[39] F. Markatopoulou et al., “ITI-CERTH participation to TRECVID 2013,” in TRECVID 2013 Workshop, 2013, pp. 12–17.

[40] C. Sun and R. Nevatia, “Large-scale web video event classification by use of fisher vectors,” *in IEEE Workshop on Applications of Computer Vision (WACV)*, 2013, pp. 15–22, Accessed: Jul. 06, 2016. [Online].

[41] R. Arandjelovic and A. Zisserman, “All about VLAD,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585, Accessed: Jul. 06, 2016. [Online].

[42] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, “Revisiting the vlad image representation,” *in Proceedings of the 21st ACM international conference on Multimedia*, Oct. 2013, pp. 653–656.

[43] G. Tolias, Y. Avrithis, and H. Jegou, “To Aggregate or Not to aggregate: Selective Match Kernels for Image Search,” i*n Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2013, pp. 1401–1408, doi: 10.1109/ICCV.2013.177.

[44] Y.-L. Boureau, J. Ponce, and Y. LeCun, “A theoretical analysis of feature pooling in visual recognition,” *in Proceedings of the 27th international conference on machine learning (ICML)*, Jun. 2010, pp. 111–118.

[DOI: 10.52547/jsdp.18.1.134]

Conference on Multimedia Retrieval, 2011, pp. 29.1-29.8.

[56] "Pretrained CNNs - MatConvNet," 2017, Accessed: Jun. 12, 2017. [Online]. Available: http://www.vlfeat.org/matconvnet/pretrained/.

[57] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27.1-27.27, 2011.

[58] "Matlab VideoUtils," SourceForge, 2015. https://sourceforge.net/projects/videoutils/ (accessed May 29, 2016).

[59] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for matlab," *in Proceedings of the 23rd Annual ACM Conference on Multimedia*, Oct. 2015, pp. 689–692.

[60] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," *in Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1469–1472.

[61] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang, "Super Fast Event Recognition in Internet Videos," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1174–1186, 2015.

[62] P. Napoletano, "Visual descriptors for content-based retrieval of remote-sensing images," *International Journal of Remote Sensing*, vol. 39, no. 5, pp. 1343–1376, 2018.

[63] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," *in Proceedings of the European Conference on Information Retrieval*, 2005, pp. 345–359.

[64] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *in Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2010, pp. 143–156.

[65] Z. Xu, Y. Yang, I. Tsang, N. Sebe, and A. G. Hauptmann, "Feature weighting via optimal thresholding for video analysis," *in Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3440–3447.

[66] A. J. Ma and P. C. Yuen, "Reduced analytic dependency modeling: Robust fusion for visual recognition," *International journal of computer vision*, vol. 109, no. 3, pp. 233–251, 2014.

[67] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang, "Robust late fusion with rank minimization," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3021–3028.

[68] I.-H. Jhuo et al., "Discovering joint audio–visual codewords for video event detection," *Machine vision and applications*, vol. 25, no. 1, pp. 33–47, 2014.

[69] D. Liu, K.-T. Lai, G. Ye, M.-S. Chen, and S.-F. Chang, "Sample-specific late fusion for visual category recognition," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 803–810.

[70] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," *in Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 167–176.

محمد سلطانیان کارشناسی، کارشناسی ارشد، و دکترای خود را در دانشگاه صنعتی شریف گذراند و هم‌اکنون استادیار گروه علوم رایانه دانشکده علوم ریاضی و رایانه در دانشگاه خوارزمی است. این کار پژوهشی مرتبط با پژوهش‌هایی است که وی در دوره دکترا انجام داده است. علایق پژوهشی ایشان شامل هوش مصنوعی و یادگیری ماشین، پردازش تصویر و ویدئو، و پردازش صحبت است.

نشانی رایانامه ایشان عبارت است از:

m.soltanian@khu.ac.ir

شاهرخ قائم‌مقامی کارشناسی و کارشناسی ارشد خود را در دانشگاه صنعتی شریف گذرانده و مدرک دکتری خود را از دانشگاه صنعتی کوئینزلند در بریزبین استرالیا اخذ کرده است. ایشان در حال حاضر عضو هیئت‌علمی پژوهشکده الکترونیک و گروه مخابرات سیستم دانشکده مهندسی برق دانشگاه صنعتی شریف است. علایق پژوهشی وی در زمینه‌های پردازش تصویر، صوت، ویدئو، و نهان‌سازی اطلاعات است. دکتر قائم‌مقامی عضو IEEE و ACM و عضو هیئت تحریریه و کمیته علمی چندین مجله و کنفرانس است.

نشانی رایانامه ایشان عبارت است از:

ghaemmag@sharif.edu

[DOI: 10.52547/jsdp.18.1.134]