بیست و دومین کنفرانس مهندسی برق ایران
دانشگاه شهید بهشتی، ۳۰ اردیبهشت ماه الی ۱ خردادماه ۱۳۹۳
The 22nd Iranian Conference on Electrical Engineering (ICEE 2014)
Shahid Beheshti University, May 20-22, 2014, Tehran, Iran

بهبود کیفیت سرویس کاربردهای جریان‌سازی ویدئو روی شبکه‌های بی‌سیم با استفاده از یادگیری تقویتی

مریم کریمی[۱]، وصال حکمی[۲]، مهدی دهقان تخت فولادی[۳]، سید مجید نورحسینی[۴]
[۱]دانشگاه صنعتی امیرکبیر، karimimaryam@aut.ac.ir
[۲]دانشگاه صنعتی امیرکبیر، vhakami@aut.ac.ir
[۳]دانشگاه صنعتی امیرکبیر، dehghan@aut.ac.ir
[۳]دانشگاه صنعتی امیرکبیر، majidnh@aut.ac.ir

چکیده - سیستم های ارتباطی حساس به تاخیر اغلب در شرایط متغییر با زمان مثل محوشدگی کانال، بار ترافیک پویا (مانند نرخ بیت متفاوت) و محتوای متغیر فعالیت می‌کنند. در چنین سیستمی نگرانی اصلی تحویل مطمئن داده به گیرنده با در نظرگرفتن تاخیر قابل تحمل است. در این مقاله انتقال بهینه نقطه به نقطه ویدئوی کدشده لایه‌ای روی کانال‌های بی‌سیم مدنظر است. راهکار ارائه شده پویایی تصادفی سیستم، ناشی از وضعیت متغیر کانال، وضعیت اشغال بافر فریم‌های ویدئو و وضعیت متغیر محتوای ویدئو را در قالب چارچوب یکپارچه بهینه‌سازی در نظر می-گیرد. هدف، بیشینه‌سازی متوسط کیفیت ویدئو دریافتی با لحاظ نمودن محدودیت متوسط تأخیر است. بنابراین، مسئله بهینه‌سازی تصادفی به عنوان یک فرایند تصمیم‌گیری مارکف فرموله‌بندی و با استفاده از یادگیری تقویتی بصورت آنلاین حل می‌شود. جهت بهبود سرعت همگرایی الگوریتم یادگیری، رویکرد تفکیک پویایی سیستم به دو مؤلفه شناخته‌شده و ناشناخته استفاده می‌شود. رویکرد مورد نظر امکان استفاده از دانش قطعی پویایی سیستم را فراهم نموده و از حجم اطلاعاتی که توسط الگوریتم یادگیری باید فراگیری شود، می‌کاهد. همچنین، با اتخاذ تصمیمات صرفا حریصانه سرعت تطبیق و کارایی الگوریتم یادگیری افزایش می‌یابد.

کلیدواژه- انتقال ویدئو، شبکه‌های بی‌سیم، فرآیند تصمیم‌گیری مارکف، یادگیری تقویتی، PDS.

۱- مقدمه

ظهور تکنولوژی شبکه‌های بی‌سیم و در دسترس بودن دوربین-های کم هزینه، کاربردهای زیادی شامل شبکه دوربینی مراقبت تا ارتباطات تصویری در عملیات امداد و نجات را در بر می‌گیرد. در چنین سناریوهایی جریانهای ویدئویی مختلفی در مناطق مختلف ساخته می شود و نیاز است که به صورت زنده کدگذاری شده و بر روی یک شبکه و بدون تاخیر انتقال داده شود. با توجه به ماهیت متغیر با زمان کانال‌های بی‌سیم و محتوای ویدئو، نیاز برای رسانه مقیاس‌پذیر که بتواند بدون کاهش کیفیت ویدئو رمزگشایی شده، نرخ بیت ویدئو را تطبیق کند، وجود دارد.

یکی از مهمترین چالش‌ها در فراهم کردن کیفیت سرویس روی یک شبکه بی‌سیم از ماهیت کانال ناشی می‌شود. قدرت سیگنال در کانال‌های بی‌سیم به دلایلی مانند فاصله، موانع و تداخل‌ها کاهش می‌یابد [۱]. این پدیده موجب اعوجاج سیگنال و خطای بسته در دریافت کننده می‌شود، بنابراین اطمینان از ارائه مطمئن بسته یک ضرورت است. با توجه به اینکه بسته‌های خطادار باید دوباره انتقال داده شوند، شبکه مجبور به نگهداری بسته‌ها در بافر تا تحویل موفق آنها به مقصد است. همچنین ورود بسته‌ها و نمونه‌گیری از داده‌های ویدئویی بصورت تصادفی انجام می‌شود در نتیجه ترافیک قابل پیش‌بینی نیست و ماهیتی تصادفی دارد. این موضوع تخصیص فضای بافر را سخت می‌کند زیرا فضای بافر در شبکه‌های بی‌سیم محدود است. از آنجایی که نرخ فریم‌های ویدئویی نسبتا بالاست و پارامترهای دوربین (مانند فاصله کانونی، موقعیت و زاویه دید) معمولا نمی‌توانند بین فریم‌ها به سرعت تغییر کنند، محتوای فریم‌های متوالی شبیه هم هستند مگر اینکه جسم معینی، با سرعت در صحنه جابجا شود. بنابراین نیازی به کدگذاری مستقل هر فریم ویدئو به عنوان یک تصویر جدید نیست و فقط تفاوت بین فریم‌های پشت سر هم، میانگین اختلاف مطلق (MAD) کدگذاری می‌شود [۲]. حال اگر لایه‌های ویدئو را درگیر کنیم اگر یک لایه نسبت به لایه دیگر متفاوت کوانتیزه شده باشد وMAD لایه اول موجود باشد،MAD لایه دوم هم قابل تخمین است. اگر تخمین MAD لایه پایه یعنی اینکه تفاضل فریمی که در لایه پایه کد می‌شود با فریم اصلی، معلوم باشد، برای لایه‌های بهبود نیازی به محاسبه MAD نیست. می-توان از MAD لایه پایه تخمین زد. بنابراین یک سیاست موفق برای انتقال نقطه به نقطه ویدئو ملزوم در نظرگرفتن تغییرات محتوای فریم، ظرفیت بافر و شرایط متغیرکانال است. کارهایی

که برای انتقال ویدئو انجام شده و تضمین محدودیت را در نظر گرفته‌اند، می‌توان به چند دسته تقسیم‌بندی کرد:

رویکرد بهینه‌سازی بین لایه‌ای: بهینه‌سازی بین لایه‌ای می‌تواند بهبود قابل توجهی در کارایی کاربر انجام دهد. هرچند اکثر الگوریتم‌های بین لایه‌ای با چندین محدودیت از جمله موارد زیر، مواجه هستند.

رویکردهای متمرکز: اکثر الگوریتم‌های بهینه‌سازی بین لایه‌ای استراتژی انتقال متمرکز دارند [۳, ۴]. به این دلیل که از معماری لایه‌ای شبکه تجاوز می‌کنند، لایه‌های مستقل توانایی برای کنترل الگوریتم‌ها و پروتکل‌های خود را از دست می‌دهند. بعلاوه اینگونه راه حل‌ها پیچیدگی خیلی بالایی دارند و سربار انتقالی که لایه‌های شرکت کننده نیاز دارند، زیاد است.

رویکردهایی با دید نزدیک‌بینانه: محدودیت رایج دیگر در الگوریتم‌های موجود این است که، بر روی ماکزیمم کردن سودمندی لحظه‌ای بدون درنظر گرفتن تاثیر عمل جاری کاربر روی کارایی دراز مدت، تمرکز دارند [۵, ۶]. در کاربردهای چندرسانه‌ای بی‌سیم، طراحی استراتژی نزدیک‌بینانه می‌تواند منجر به تباهی غیر قابل قبول کیفیت چندرسانه‌ای در دراز مدت به دلیل خصوصیات ناهمگن ترافیک رسانه، شود [۶, ۷].

رویکرد بهینه‌سازی آفلاین: در این نوع بهینه‌سازی، فرض می‌شود که پارامترهای آماری سیستم مانند احتمالات گذار و شرایط تصادفی از قبل معلوم هستند بر این اساس راهکارها و الگوریتم-هایی داده می‌شود تا بتوان سیاست بهینه را محاسبه کرد. در عمل این اطلاعات در دسترس نیست و باید در حین اجرای سیستم یادگیری شوند.

دسته دیگر ازکارهای انجام شده وجود دارند که تغییرات محتوای فریم، ظرفیت بافر و شرایط متغیرکانال را بصورت یکجا درنظر نگرفته‌اند:

مطالعه [۸] شرایط متغیر کانال و نرخ ورود بسته‌ها به بافر را به عنوان حالت سیستم در فرایند تصمیم‌گیری مارکف در نظر گرفته ولی تغییرات محتوای ویدئو در آن دیده نشده است. در [۹-۱۱] ویژگی متغیر با زمان کانال درنظر گرفته نشده است. در [۱۰] مقاله نرخ ورود بسته به بافر ثابت و سناریو چند نقطه به نقطه است. حل آن هم با استفاده از نظریه بازی‌ها صورت گرفته است. در الگوریتم یادگیری که استفاده شده، سرعت همگرایی پایین است در واقع دسته‌ای از اطلاعات سیستم وجود دارند که نیازی نیست در چرخه یادگیری قرار بگیرند چون پویایی آن‌ها شناخته شده است. می‌توان از تکنیکی برای یادگیری استفاده کرد که پویایی‌های شناخته شده را از پویایی‌های ناشناخته سیستم، تفکیک می‌کند.

در این مقاله انتقال نقطه به نقطه بهینه داده حساس به تاخیر چندرسانه‌ای بر روی کانال‌های بی‌سیم مدنظر است. درحال حاضر چارچوب یکنواختی برای بهینه سازی تصادفی کیفیت ویدئو با در نظر گرفتن محدودیت تاخیر و نیز شرایط تصادفی و نامعلوم محتوای ویدئو، بافر و کانال وجود ندارد. در اینجا ارائه این چارچوب مدنظر است و مسئله به عنوان یک فرایند تصمیم‌گیری مارکف فرموله‌بندی شده و با استفاده از یادگیری تقویتی به صورت آنلاین حل می‌شود. نوآوری‌های مقاله عبارت است از (*i*) حل مساله بدون نیاز به پارامترهای تصادفی شرایط کانال و تغییرات محتوای ویدئو (*ii*) تفکیک پارامترهای معلوم از پارامترهای ناشناخته سیستم؛ یعنی به کارگیری الگوریتم یادگیریPDS. روش ارائه شده نیاز به اطلاعات کمتری نسبت به الگوریتم‌های یادگیری تقویتی سنتی برای یادگیری، دارد و نیازی به کشف عمل، که شدیدا سرعت تطبیق الگوریتم‌های یادگیری تقویتی سنتی را کاهش می‌دهد، ندارد.

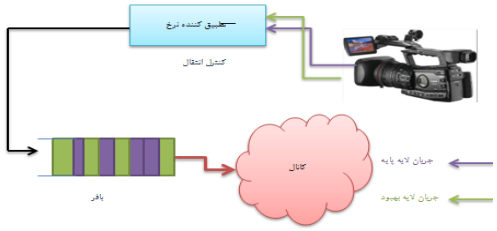
مقاله طبق زیر سازماندهی شده است: در قسمت ۲ از مقاله به تعریف مساله و فرمول‌بندی آن پرداخته می‌شود قسمت ۳ راه‌حل مطرح شده و سیاست بهینه را شرح خواهد داد. در قسمت ۴ شبیه‌سازی الگوریتم پیشنهادی ارائه می‌شود و در نهایت هم نتیجه‌گیری و کارهای آتی بیان خواهد شد.

۲- تعریف مساله و فرمول‌بندی

داده از طریق کانال بی‌سیم به سمت فرستنده ارسال می‌شود. سیستم بصورت نقطه به نقطه در نظر گرفته شده است. کاربر دارای یک بافر و یک تطبیق کننده نرخ جهت انتقال ویدئو کدشده لایه‌ای (SVC) است که خود از یک لایه پایه و چند لایه بهبود تشکیل شده است (شکل۱). بعد از ورود تصادفی داده از دوربین به بافر باید داده‌ها به سمت گیرنده دیگر، از طریق کانال بی‌سیم ارسال شوند. ارسال روی کانالی انجام می‌شود که در معرض تغییرات تصادفی است. فرض کنید کاربر جریان ویدئوی خود را با پارامتر کوانتیزه q_l کدگذاری می‌کند بطوریکه $l=0$ و $l>=1$ به ترتیب متناظر با لایه پایه و هر لایه بهبود هستند. این کدگذاری منجر به نوسان نرخ بیت و اعوجاج به عنوان تابعی از پیچیدگی تصویر و سطح تغییرات حرکت، می‌شود. نوسان ویدئو را می‌توان MAD سیگنال‌های پیش‌بینی شده، درنظر گرفت. زیرا MAD تفاوت بین تصویر اصلی کدگذاری نشده و پیش‌بینی شده را کمّی می‌کند. تصمیم‌گیری بر روی ارسال تعداد فریم‌های

ویدئویی است که باید با توجه به وضعیت کانال و ظرفیت بافر انجام شود.

حال مساله بصورتCMDP و با چندتایی حالت سیستم، عمل کنترلی، پاداش انتقال و محدودیت تاخیر فرموله‌بندی می‌شود.



شکل ۱: مدل سیستم

حالت سیستم: فرض می‌شود b^n معرف وضعیت اشغال بافر در زمان n و $b^n \in \{0,1,\dots,B\}$ است. B حداکثر تعداد فریم‌های ویدئویی کدگذاری شده که می‌تواند در بافر ذخیره شود، می‌باشد. کیفیت کانال با h^n نشان داده شده و با کوانتیزه کردن شاخص کیفیت کانال (نسبت توان سیگنال به نویز یا SNR)، فضای حالت کانال را می‌توان با $h^n \in \{1,2,\dots,H\}$ نمایش داد که تغییرات آن را طی گام‌های متوالی می‌توان بصورت زنجیره مارکف مدل کرد [۱۲]. فرض می‌شود که m^n، MAD فریم در زمان n است. در [۱۳]، MAD بین دو فریم پشت سرهم به عنوان یک فرآیند مارکف مدلسازی می‌شود. فضای حالت تغییرات ویدئو با $\mathcal{M} = \{0,1,\dots,M\}$ نشان داده می‌شود، بازه عددی m^n کوانتیزه شده و حالت‌های ویدئو یک فرآیند مارکف را تشکیل می‌دهند. در نهایت حالت سیستم در زمان n، ترکیبی از حالت کانال، حالت ویدئو و حالت بافر است؛ یعنی $s^n=\{b^n,h^n,m^n\}$.

عمل کنترلی: عمل a^n توسط کاربر در بازه زمانی n انجام می‌شود که عبارتست از تعداد فریم‌های ویدئویی کد شده که برای انتقال از بافر خارج می‌شوند. هر بسته ورودی به بافر دارای l لایه می‌باشد.

پاداش انتقال: مقادیر QP فریم‌های ویدئویی کدشده بلادرنگ در لایه پایه و بهبود تغییر می‌کنند. فرض می‌شود که $q_l, q_{l'} \in \{0,1,\dots,50\}$به ترتیب مقادیر QP فعلی و قبلی هستند. در معادله (۱)، MAD باقیمانده l می‌تواند با استفاده از MAD باقیمانده l' محاسبه شود:

$$\widetilde{m}_l = \widetilde{m}_l 2^{\varsigma\,(q_l - q_l)} \qquad (۱)$$

ς پارامتر مدل و بطور معمول برای اکثر توالی‌ها، ۰,۰۷ است. یک مدل اعوجاج که حداکثر نسبت توان سیگنال به نویز تصویر کدشده را تخمین می‌زند، در معادله (۲) آمده است [۱۴]:

$$\delta(l,\widetilde{m}_l) = v_1\, log_{10}((\widetilde{m}_l)^u + 1).\ q + v_2 = \quad (۲)$$
$$v_1\, log_{10}\left(\left(\widetilde{m}_l 2^{\varsigma\,(q_l - q_l)}\right)^u + 1\right).\ q + v_2,$$

که در آن u فاکتور توان با مقدار $5/6$ است. v_1 و v_2 عبارتند از پارامترهای مدل وابسته به توالی و مقدار آنها به ترتیب ۰,۵۲ و۴۷ است. $r(s^n,a^n)$ پاداش انتقال کاربر در زمان n و در واقع PSNR ویدئویی مورد انتظار ناشی از انتقال لایه‌های ویدئویی l_{a^n} از فریم‌های ویدئویی ذخیره شده در بافر کاربر است و l_{a^n} تعداد لایه‌های فریم‌های انتقالی با انجام عمل a^n است. $\delta(l,m^n)$، PSNR حاصل از کدگشایی همه فریم‌ها تا لایه l کاربر در زمان n و در حالت ویدئو m^n می‌باشد. $p_l(h^n)$ نرخ خطای بسته برای کاربر در l‌امین لایه، در حین انتقال است. معادله (۳) پاداش کاربر (PSNR) را نشان می‌دهد [۱۵].

$$r(s^n,a^n) = a^n.\Big[\big(1-p_0(h^n)\big).\ \delta(0,m^n) + \quad (۳)$$
$$\textstyle\sum_{l=1}^{l_{a^n}-1}\prod_{j=0}^{l}\big(1-p_j(h^n)\big).\big[\delta(l,m^n) -$$
$$\delta(l-1,m^n)\big]\Big]$$

محدودیت تاخیر: کاربر یک کیفیت سرویس لحظه‌ای به لحاظ محدودیت تاخیر فریم‌های داخل بافر دارد و با $d(s^n,a^n)$ نشان داده می‌شود. اگر تاخیر به عنوان معیار محدودیت درنظر گرفته شود، در نتیجه $d(s^n,a^n)$ تابعی از حالت بافر b^n کاربر خواهد بود. درگام زمانی n تعداد فریم‌های باقیمانده در بافر طبق فرمول (۴) قابل محاسبه است:

$$d(s^n,a^n) = (b^n - a^n + f_{in}) \quad (۴)$$

f_{in} بیانگر تعداد فریم‌های ورودی به بافر است. d را می‌توان به عنوان شاخصی از تاخیر لحظه‌ای، درنظر گرفت. هدف، یافتن سیاست بهینه ارسال فریم‌های ویدئو π^* به منظور بیشینه‌سازی متوسط تخفیفی PSNR و تضمین محدودیت متوسط تخفیفی تاخیر است و طبق فرمول زیر بیان کرده‌ایم:

$$\pi^* \in argmax\ PSNR(\pi) = \quad (۵)$$
$$E_{s^n}[\textstyle\sum_{n=1}^{\infty}\gamma^{n-1}.\ psnr(s,a^n)|\ \pi]$$

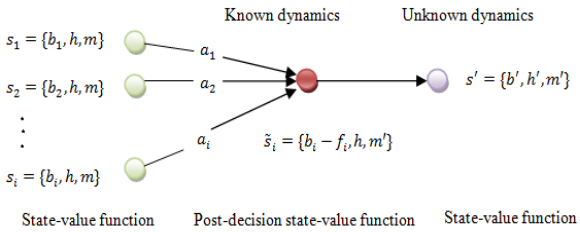
که $0 \le \gamma < 1$ و π به ترتیب فاکتور کاهش و سیاست انتقال هستند. با استفاده از رابطه (۶) محدودیت تاخیر را بیان می‌کنیم:

$$D(\pi) = E_{s^n}[\textstyle\sum_{n=1}^{\infty}\gamma^{n-1}.\ d(s,a^n)|\ \pi] \quad (۶)$$

۳- یادگیری سیاست بهینه انتقال ویدئو

به دلیل اینکه احتمالات گذار، پاداش،PSNR و... به تغییرات وابسته‌اند و از قبل معلوم نمی‌باشند. بنابراین برای حل مساله به

دنبال یادگیری هستیم. تعدادی از یادگیری‌های بدون مدل مانند Q-learning وجود دارند اما مدل دیگر از یادگیری PDS است که پویایی‌های شناخته شده و ناشناخته را از هم تفکیک می‌کند که این تکنیک باعث می‌شود که اطلاعات معلوم را بی-جهت وارد عملیات یادگیری نکنیم تا سرعت انجام الگوریتم بالاتر برود [۱۶, ۱۷]. از طرفی به عمل اکتشاف که زمان زیادی طول می‌کشد، نیازی نیست. همانطور که در شکل۲ آمده است، PDS حالت سیستم بعد از انجام پویایی‌های معلوم سیستم، اما قبل از اینکه پویایی‌های نامعلوم اتفاق بیفتد، است. نکته دیگر اینکه از چندین حالت-عمل متفاوت می‌توان به یک حالت PDS یکسان رسید. PDS با $\tilde{s}$ نشان داده می‌شود. در این مقاله مساله به کمک PDS حل خواهد شد.



شکل ۲: رابطه بین جفت عمل-حالت در زمان n، PDS در زمان n و حالت در زمان n+1، چند حالت-عمل مختلف می‌توانند PDS یکسان داشته باشند.

PDS بافر، حالت بافر بعد از انتقال فریم‌های موجود اما قبل از رسیدن فریم‌های جدید است. اما PDS کانال همان حالت کانال در بازه زمانی n است. به عبارت دیگر PDS از همه اطلاعات معلوم درباره انتقال از حالت s^n به s^{n+1} بعد از انجام عمل a^n بهره می‌گیرد. در حالی که حالت بعدی از اطلاعات نامعلومی که در PDS دخالت داده نشده‌اند، تشکیل می‌شود (یعنی تعداد فریم‌های رسیده، l^n و حالت بعدی کانال h^{n+1} بهره می‌گیرد. حالت بافر را در بازه زمانی $n+1$ بر اساس PDS بافر در بازه زمانی n به صورت فرمول (۷) دوباره نویسی می‌کنیم:

$$b^{n+1} = \tilde{b}^n + l^n \qquad (۷)$$

الگوریتم یادگیری PDS طبق مراحل زیر انجام می‌شود:

آغاز: در زمان $n = 0$، تابع PDS مقدار دهی اولیه می‌شود.

انجام عمل حریصانه: در زمان n، عمل حریصانه را طوری انتخاب می‌کنیم که معادله (۸) را ماکزیمم کند.

$$a_n = \underset{a\epsilon A}{argmax}\{reward_k(s_n, a) - \lambda_n * (b_n - a_n - \Lambda) + \tilde{V}_n(\tilde{s})\} \qquad (۸)$$

ارزیابی تابع مقدار-حالت: مقدار حالت s^{n+1} را از معادله زیر بدست می‌آوریم:

$$V_n(s_{n+1}) = \underset{a\epsilon A}{max}\{reward_k(s_{n+1}, a) - \lambda_n * (b_{n+1} - a_n - \Lambda) + \tilde{V}_n(\tilde{s})\} \qquad (۹)$$

طبق این معادله ماکزیمم مقداری که یک عمل باعث آن می-شود، به دست می‌آید، Λ مقدار محدودیت و λ، ضریب لاگرانژ است.

سپس می‌توان تابع PDS را با استفاده از رابطه (۱۰) به روز رسانی کرد:

$$\tilde{V}_{n+1}(\tilde{s}_n) = (1 - \alpha_n)\tilde{V}_n(\tilde{s}_n) + \alpha_n[\lambda_n * (\tilde{b}_n + pois\ srn(\Delta) - \Lambda) + \gamma V_n(s_{n+1})] \qquad (۱۰)$$

که $reward_u$ را از رابطه (۱۱) محاسبه می‌کنیم:

$$reward_u = \lambda_n(b_n + l - \Delta) \qquad (۱۱)$$

l برابر تعداد فریم‌های ورودی است که بصورت تصادفی وارد بافر می‌شوند. علائم u و k نشانگر ویژگی‌های نامعلوم و معلوم سیستم هستند.

مقدار بهینه ضریب لاگرانژ λ که وابسته به محدودیت تاخیر میانگین،کاهش یافته است را بطور آنلاین طبق رابطه (۱۲) یاد می‌گیریم:

$$\lambda_{n+1} = \lambda_n + \beta_n(b_n - (1 - \gamma)\delta) \qquad (۱۲)$$

β^n نرخ یادگیری متغیر با زمان با ویژگی‌های مشابه α^n می‌باشد. b^n مقدار بافر در زمان n است. اصطلاح $(1-\gamma)\delta$ تاخیر کاهشی را به محدودیت تبدیل می‌کند. شرایط زیر نیز باید بوسیله β^n و α^n برای اطمینان از همگرایی رابطه (۱۳) به μ بهینه برآورده شود [۷].

$$\sum_{n=0}^{\infty}(\alpha^n + \beta^n) < \infty \ and \ \lim_{n\to\infty} \frac{\beta^n}{\alpha^n} \to 0. \qquad (۱۳)$$

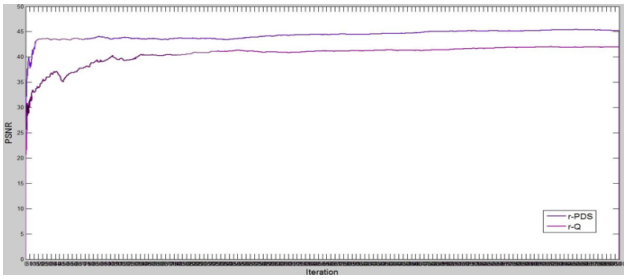
چندین دلیل برای نشان دادن اینکه یادگیری PDS اطلاعات کاراتری نسبت به Q-learning استفاده می‌کند، وجود دارد. اول اینکه یادگیری PDS از اطلاعات مختصر سیستم هم بهره می‌برد تا اطلاعات کمتری برای یادگیری نیاز داشته باشد. استفاده از $reward_k$ در معادلات (۸) و(۹) مؤید این موضوع است. دوم، به روز رسانی PDS با استفاده از معادله (۱۰) اطلاعاتی درباره تابع مقدار-حالت تعداد زیادی حالت را فراهم می‌کند. این مهم از مقدار PDS موردانتظار در سمت راست معادله (۹) معلوم است. سوم اینکه، به دلیل بستگی نداشتن ویژگی $reward_u$ به عمل،

نیازی به اکتشاف تصادفی به منظور پیدا کردن عمل بهینه در هر حالت وجود ندارد. این بدین معناست که آخرین تابع مقدار تخمین شده را می‌توان همیشه به کار برد و نیازی به تست کردن عمل غیر حریصانه نیست.

۴- شبیه‌سازی

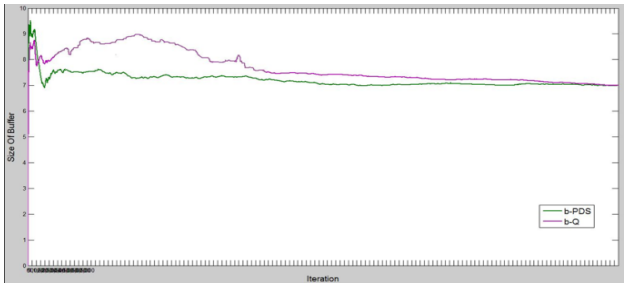
سناریوی ارزیابی شده بصورت نقطه به نقطه می‌باشد. هر فریم ویدئویی از یک لایه پایه و دو لایه بهبود تشکیل و ورود فریم به بافر بصورت پواسون و حداکثر سه فریم در هر بازه زمانی (۱میلی ثانیه) است. ظرفیت بافر ویدئو۱۰ فریم در نظرگرفته و کانال به ۸ محدوده احتمال مساوی تفکیک شده است. محدودیت تعریف شده بر روی تعداد فریم‌های موجود در بافر که در واقع متوسط قید بافر است، ۷ فریم تعریف شده تا تاخیر از حدی بیشتر نشود. فاکتور کاهش γ، برابر ۰٫۸۵ و توابع α_n و β_n به ترتیب بصورت $1/n^{0.55}$ و $1/n$ تنظیم شده که در آن‌ها n شاخص تکرار و برابر ۷۰۰۰ می‌باشد. معیارهای شبیه‌سازی که جهت ارزیابی کارایی الگوریتم پیشنهادی استفاده می‌شوند، عبارتند از:

شکل۳ میزان PSNR در برآوردن کیفیت ویدئو را در دو الگوریتم Q-learning و PDS نشان می‌دهد. همانطور که در نمودار دیده می‌شود مقدار پاداش الگوریتم PDS از Q-learning بیشتر بوده از طرفی نمودار آن سریعتر همگرا شده است.



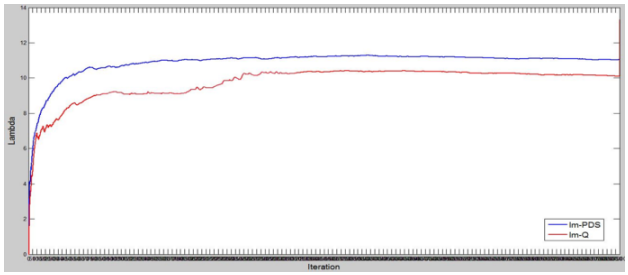
شکل ۳:مقایسه PSNR در دو الگوریتم یادگیری PDS وQ-learning

شکل۴ پرشدگی بافر در طول زمان اجرای الگوریتم را نشان می‌دهد و به این مهم اشاره دارد که در هر دو الگوریتم سعی در رسیدن به محدودیت تاخیر تعیین شده است و اینکه از آن تجاوز نشود. مقدار محدودیت تاخیر همان ۷ فریم می‌باشد. همانطور که دیده می‌شود سرعت همگرایی PDS بیشتر است.



شکل ۴: برآورده کردن محدودیت تاخیر و همگرایی سریع‌تر PDS نسبت به Q-learning

همگرا شدن ضریب لاگرانژ برای رسیدن به جواب بهینه لازم است. در شکل ۵ همگرایی در هر دو الگوریتم وجود دارد و همانطور که انتظار می‌رود سرعت همگرایی در PDS بالاتر است.



شکل ۵: همگرایی ضریب لاگرانژ. شیب الگوریتم PDS از Q-learning بیشتر است.

۵- نتیجه‌گیری

در این مقاله، با درنظر گرفتن شرایط متغیر کانال، ویدئو و وضعیت بافر مسأله با استفاده از چارچوب تصمیم‌گیری مارکف فرموله‌بندی شد. در ادامه بر اساس مفهوم PDS، پویایی‌های سیستم به پویایی‌های شناخته شده و ناشناخته تجزیه شدند. بنابراین فرآیند تصمیم‌گیری مارکف دوباره فرموله‌بندی شد که اطلاعات معلوم سیستم را بی‌جهت در فرآیند یادگیری دخالت ندهد و بنابراین نسبت به الگوریتم یادگیری تقویتی سنتی نیاز به فراگیری اطلاعات کمتری دارد. همچنین مزیت دیگر رویکرد پیشنهادی این است که نیازی به اقدام اکتشافی، که سرعت عملکرد الگوریتم یادگیری تقویتی سنتی را شدیدا کاهش می‌دهد، ندارد. این روش یادگیری به دلایل گفته شده نسبت به الگوریتم‌های یادگیری سنتی سریعتر به مقدار بهینه همگرا می‌شود. ارزیابی و نتایج حاصل از شبیه‌سازی نشان‌دهنده کارایی بالای الگوریتم پیشنهادی PDS در بهبود کیفیت ویدئو است. برای کارهای آتی می‌توان محدودیت بیشتر علاوه بر تاخیر را برای بهتر شدن کیفیت در نظر گرفت، و یا مساله را بصورت چند نقطه به نقطه حل کرد و یا اینکه از روش پیشنهادی برای وقتی که فرآیند ورود به بافر لزوما $i.i.d$ نیست یا تغییرات محتوای

ویدئو غیر مارکفی است، استفاده کرد.

مراجع

[1] B. Sklar, "Rayleigh fading channels in mobile digital communication systems. I. Characterization," *Communications Magazine, IEEE,* vol. 35, pp. 90-100, 1997.

[2] Z.-N. Li and M. S. Drew, *Fundamentals of multimedia, ISBN: 0130618721* vol. 7458: Pearson Education, Inc., Upper Saddle River, NJ, 2004.

[3] M. van Der Schaar, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *Wireless Communications, IEEE,* vol. 12, pp. 50-58, 2005.

[4] V. Srivastava and M. Motani, "Cross-layer design: a survey and the road ahead," *Communications Magazine, IEEE,* vol. 43, pp. 112-119, 2005.

[5] F. Zhai, Y. Eisenberg, and A. K. Katsaggelos, "Joint source-channel coding for video communications," *Handbook of Image and Video Processing,* 2005.

[6] R. Hamzaoui, V. Stankovic, and Z. Xiong, "Optimized error protection of scalable image bit streams [advances in joint source-channel coding for images]," *Signal Processing Magazine, IEEE,* vol. 22, pp. 91-107, 2005.

[7] N. Salodkar, A. Bhorkar, A. Karandikar, and V. S. Borkar, "An on-line learning algorithm for energy efficient delay constrained scheduling over a fading channel," *Selected Areas in Communications, IEEE Journal on,* vol. 26, pp. 732-742, 2008.

[8] H.-P. Shiang and M. van der Schaar, "Online learning in autonomic multi-hop wireless networks for transmitting mission-critical applications," *Selected Areas in Communications, IEEE Journal on,* vol. 28, pp. 728-741, 2010.

[9] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," *Multimedia, IEEE Transactions on,* vol. 8, pp. 390-404, 2006.

[10] H. Wang and A. Ortega, "Rate-distortion optimized scheduling for redundant video representations," *Image Processing, IEEE Transactions on,* vol. 18, pp. 225-240, 2009.

[11] C .De Vleeschouwer and P. Frossard, "Dependent packet transmission policies in rate-distortion optimized media scheduling," *Multimedia, IEEE Transactions on,* vol. 9, pp. 1241-1258, 2007.

[12] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," *Vehicular Technology, IEEE Transactions on,* vol. 44, pp. 163-171, 1995.

[13] X. Zhu and B. Girod, "Analysis of multi-user congestion control for video streaming over wireless networks," in *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 869-872, 2006.

[14] H. Mansour, P. Nasiopoulos, and V. Krishnamurthy, "Real-time joint rate and protection allocation for multi-user scalable video streaming," in *Personal, Indoor and Mobile Radio Communications, 2008 .PIMRC 2008. IEEE 19th International Symposium on*, pp. 1-5, 2008.

[15] J. W. Huang, H. Mansour, and V. Krishnamurthy, "A dynamical games approach to transmission-rate adaptation in multimedia WLAN," *Signal Processing, IEEE Transactions on,* vol. 58, pp. 3646-3635, 2010.

[16] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning,* vol. 16, pp. 185-202, 1994.

[17] D. P. Bertsekas, "Dynamic programming and optimal control 3rd edition, volume II," *Belmont, MA: Athena Scientific,* 2011.