تشخیص بصری گفتار با استفاده از تحلیل مکان-زمانی گرادیان

علی جعفری شش پلی[۱]؛ علی نادیان قمشه[۲]

۱- پژوهشکده فضای مجازی - دانشگاه شهید بهشتی - تهران - ایران - ali.jafari1@mail.sbu.ac.ir
۲- پژوهشکده فضای مجازی - دانشگاه شهید بهشتی - تهران - ایران - a_nadian@pmail.sbu.ac.ir

چکیده: استفاده از اطلاعات بینایی برای تشخیص گفتار، راه‌کاری مهم در عدم حضور اطلاعات صوتی است. در این مقاله، روشی برای تشخیص گفتار به کمک اطلاعات بینایی با توصیف تغییرات مکانی-زمانی ناحیه لب ارائه شده است. برای توصیف تغییرات از گرادیان تصویر استفاده شد. در روش پیشنهادی، پس از تشخیص ناحیه لب و استخراج نقاط کلیدی، گرادیان در نواحی مربوط به نقاط کلیدی به‌عنوان اطلاعات مکانی مورداستفاده قرار گرفت. برای توصیف نواحی کلیدی لب در طول بیان یک عبارت، نمودار فراوانی ۳ بعدی گرادیان‌ها و تخمین مسیر تغییرات نواحی کلیدی در طول ویدیو استفاده شدند. تمرکز اصلی این تحقیق، ارائه توصیفی مناسب از گفتار است. به همین منظور، از دسته‌بندهای متفاوتی برای تشخیص گفتار به کمک ویژگی‌های استخراج شده استفاده شد تا دسته‌بند مناسب‌تر مورد استفاده قرار گیرد. برای ارزیابی روش پیشنهادی از بانک داده MIRACL-VC1 استفاده شد و نتایج به دست آمده با روش‌های پیشین برای تشخیص گفتار مقایسه شدند. نتایج نشان داد روش پیشنهادی در حدود ۱۱ تا ۱۷ درصد بهبودی داشته است.

واژه‌های کلیدی: تشخیص بصری گفتار، گرادیان زمانی و مکانی، تطبیق منحنی، ویژگی‌های ظاهری، ویژگی‌های حرکتی.

Visual Speech Recognition using Spatial-Temporal Gradient Analysis

Ali Jafari-Sheshpoli[1]; Ali Nadian-Ghomsheh[2]

1- Cyber space research inst., Shahid Beheshti University, Tehran, Iran, Email: ali.jafari1@mail.sbu.ac.ir
2- Cyber space research inst., Shahid Beheshti University, Tehran, Iran, Email: a_nadian@pmail.sbu.ac.ir

Abstract: The use of visual information for voice recognition is an important solution in the absence of audio information. This paper presents a method for speech recognition using visual information by describing spatial-temporal changes in the lobe of the lips. The gradient of the image was used for feature extraction. In the proposed method, after lobe area detection and extraction of key points, the gradient was extracted to describe the spatial information of the key points. To describe the key areas of the lip during speaking, the 3D histogram of gradients path curve fitting was used. The main focus of this research was to provide an adequate description of speech. For this purpose, different classifiers were tested and the best one was recognized. To evaluate the proposed method, the MIRACL-VC1 database was used and the results were compared with previous methods for speech recognition which had an improvement about 11 to 17 percent.

Keywords: Visual speech recognition, temporal and spatial gradient, curve fitting, local features, motion features.

تاریخ ارسال مقاله: ۱۳۹۷/۰۵/۱۱
تاریخ اصلاح مقاله: ۱۳۹۷/۱۱/۰۶و ۱۳۹۸/۰۱/۱۶
تاریخ پذیرش مقاله: ۱۳۹۸/۰۳/۱۲
نام نویسنده مسئول: علی نادیان قمشه
نشانی نویسنده مسئول: ایران - تهران - ولنجک - دانشگاه شهید بهشتی – پژوهشکده فضای مجازی.

۱- مقدمه

گفتار به‌عنوان یکی از مؤثرترین و طبیعی‌ترین ابزار برقراری ارتباط بین انسان‌ها شناخته می‌شود. متأسفانه افراد ناشنوا و کم‌شنوا نمی‌توانند از این شکل طبیعی ارتباط استفاده کنند. لب‌خوانی یا تشخیص بصری گفتار، روشی برای تشخیص گفتار با استفاده از حرکات لب، صورت و زبان، در عدم حضور صدا است.

لب‌خوانی شامل دو مرحله‌ی استخراج ویژگی و تشخیص گفتار است. ویژگی‌هایی که در پژوهش‌های پیشین برای تشخیص گفتار استفاده‌شده‌اند عبارت‌اند از: ویژگی‌های مبتنی بر شکل، ویژگی‌های مبتنی بر ظاهر و ویژگی های مبتنی بر حرکت لب ها[۱]. توصیف مناسب و بازنمایی ناحیه لب‌ها تأثیر بسزایی در دقت تشخیص لب خوانی دارد.

ویژگی‌های مبتنی بر شکل، از شکل لب‌ها، اطلاعات هندسی مانند ارتفاع و عرض دهان، مساحت و محیط دهان و ... را استخراج می‌کنند. در این روش‌ها، ابتدا مرز لب گوینده از فریم‌ها تشخیص داده می‌شود و سپس یک مدل احتمالی یا پارامتری از مرز لب به دست می‌آید. در مرحله، بعد پارامترهای مدل لب به‌عنوان ویژگی‌های شکل لب استفاده می‌شوند. روش‌های متداول در این دسته شامل مدل مرز فعال (ACM)[۱]، مدل شکل فعال (ASM)[۲] [۲]، مدل ظاهر فعال[۳] (AAM) [۳-۷] و مدل مار[۴] [۸، ۹] می‌باشند. در مدل ظاهر فعال یک مدل آماری از شکل و ظاهر لب‌ها به یک تصویر جدید تطبیق داده می شود[۱۰]. در مدل مار، منحنی‌های انعطاف‌پذیری به‌صورت پویا به مرزهای لب تطبیق داده می‌شود[۶]. این سیستم شامل مجموعه‌ای از نقاط به‌هم‌پیوسته است که مکان نقاط آن با تغییر لب کنترل می‌شود. تعیین شیء در تصویر از طریق مرز فعال یک فرایند تعاملی است. کاربر باید مرز اولیه‌ای که تقریباً مشابه شکل لب است را تخمین بزند.

این ویژگی‌ها قادر به توصیف شکل‌های مختلف لب هستند، اما لازم به ذکر است که مرحله آموزش این دسته از روش‌ها به علت دستی بودن، بسیار زمان‌بر است[۱].

ویژگی‌های مبتنی بر ظاهر مستقیماً از یک ناحیه مستطیلی که ناحیه دهان را مشخص می‌کند استخراج می‌شوند. برخی از ویژگی‌هایی که در این زمینه استفاده‌شده‌اند عبارت‌اند از: روش‌های الگوی محلی دودویی[۵] (LBP) [۱۱، ۱۲]، تبدیل کسینوسی گسسته[۶] (DCT) [۱۲]، هیستوگرام گراد یان زاو یه‌ای[۷] (HOG) [۱، ۱۱] و تبدیل موجک گسسته[۸] (DWT) [۱۳-۱۵]. ویژگی LBP یکی دیگر از اپراتورها برای توصیف بافت است که پیکسل‌های منطقه مورد نظر را با آستانه گذاری همسایگان هر پیکسل نشان می‌دهد و نتیجه را به‌عنوان یک عدد دودویی در نظر می‌گیرد[۱۶]. DCT یک ویژگی قدرتمند برای توصیف یک تصویر در دامنه فرکانس است و تصویر را به‌صورت مجموعی از توابع کسینوسی که در فرکانس‌های مختلف نوسان دارند، بیان می‌کند. توصیفگر HOG ظاهر و شکل محلی شیء در تصویر را با توزیع شدت شیب یا جهت لبه‌ها توصیف می‌کند[۱۲]. تبدیل موجک یک تجزیه چند سطحی از تصویر ورودی انجام می‌دهد و تصویر را به چند زیرمجموعه تقسیم می‌کند که در هر سطح اطلاعات فرکانسی-مکانی مربوط ذخیره می‌شود[۱۵]. تبدیل موجک گسسته یکی از تبدیلات قدرتمند برای توصیف آماری تغییرات روشنایی سطوح است که در تحقیقات مختلفی در حوزه پردازش تصویر بکار گرفته شده است[۱۷].

ویژگی‌های مبتنی بر شکل، اطلاعات مکانی را استخراج می‌کنند و در توصیف زمانی ویژگی گفتاری ناتوان‌اند. درنتیجه، در کاربردهایی مانند لب‌خوانی که حاوی اطلاعات و ویژگی‌های زمانی مفیدی هستند، نتایج مطلوبی به دست نمی‌آید.

ویژگی‌های دسته سوم حرکت لب‌ها را دنبال می‌کنند و تغییرات زمانی را در نظر می‌گیرند. در این روش، ویژگی‌های زمانی استخراج می‌شوند و تغییرات حرکتی لب‌ها در بین فریم‌ها را توصیف می‌کنند. از ویژگی‌های استفاده‌شده در این دسته می‌توان از هیستوگرام حرکت لبه‌ها[۹] (MBH) [۱]، الگوی محلی دودویی – سه صفحه متعامد (LBP-TOP)[۱۰] [۱۸] و تاریخچه حرکت تصویر[۱۱] (MHI) [۱۹] نام برد. MBH حرکت نسبی بین پیکسل‌های ناحیه مورد نظر را بر اساس مشتقات افقی و عمودی تصویر جریان نوری[۱۲]، کدگذاری می‌کند. ازآنجایی‌که MBH نشان‌دهنده گرادیان جریان نوری است، حرکت دوربین تأثیری در جریان حرکتی نداشته و اطلاعات مربوط به تغییرات در مرزهای حرکتی حفظ می‌شود[۱]. در روش LBP-TOP برای هر پیکسل از تصویر یک همسایگی در صفحه‌های XY,XT و YT در نظر گرفته می‌شود و برای هر صفحه، ویژگی LBP استخراج می‌شود[۱۸]. تاریخچه حرکت تصویر، یک الگو از تصویر است که چگونگی تغییر یک الگوی خاص در طول زمان را نشان می‌دهد. ویژگی MHI، اطلاعات حرکتی را به یک تصویر الگو که شدت آن تابعی از زمان حرکت است، تصویر می‌کند؛ بنابراین، شدت پیکسل MHI، تابعی از تاریخ حرکت در آن مکان است، جایی که مقادیر روشن‌تر مربوط به یک حرکت اخیر است[۱۹].

ارزیابی سیستم‌های تشخیص گفتار مبتنی بر بینایی ماشین عبارت‌اند از: آزمون مستقل از گوینده و آزمون وابسته به گوینده[۱]. در آزمون مستقل از گوینده، داده‌های یک گوینده به‌عنوان داده‌های آزمون و داده‌های گویندگان باقیمانده برای مرحله آموزش مورد استفاده قرار می‌گیرد. در آزمون وابسته به گوینده، برای هر یک از گویندگان در مجموعه داده‌ها، دو ویدیو برای آزمون و بقیه برای آموزش استفاده می شنود. با وجود اهمیت موضوع تشخیص گفتار، نتایج به‌دست‌آمده دقت مطلوبی ندارند و تحقیقات بیشتری برای افزایش دقت تشخیص گفتار به کمک اطلاعات بصری مورد نیاز است.

در تحقیق حاضر، هدف ارائه روشی برای تشخیص گفتار بر اساس توصیف تغییرات مکانی-زمانی لب‌ها است. در فریم‌های خاصی از یک عبارت، لب‌ها فرم خاصی دارند و می‌توان با توصیف لب در هر فریم و تشخیص فریم‌های مشابه، عبارت‌های مشابه را پیدا کرد.

همچنین انتظار می‌رود یک ناحیه مشخص از لب در طول بیان عبارت‌های مشابه، فرم خاصی از تغییرات را دنبال کند؛ بنابراین با توصیف نحوه تغییرات نواحی لب در طول زمان می‌توان عبارت‌های مشابه را تشخیص داد. در روش پیشنهادی با استفاده از گرادیان و تخمین تغییرات شکل ناحیه لب در طول زمان اقدام به تشخیص گفتار شد. در این پژوهش گرادیان در نقاط کلیدی لب و گرادیان در سه جهت برای توصیف ناحیه لب استفاده شد. همچنین برازش منحنی بر روی نقاط کلیدی لب در طول زمان برای دنبال کردن تغییرات استفاده شد. برای ارزیابی روش پیشنهادی از بانک داده MIRACL-VC1 استفاده شد. این بانک داده شامل ۱۰ کلمه (۱۵ نفر × ۱۰ حرف × ۱۰ تکرار) و ۱۰ عبارت (۱۵ نفر × ۱۰ عبارات × ۱۰ تکرار) است [۲۰]. با توجه به اینکه تمرکز اصلی روش پیشنهادی در این پژوهش، ارائه روشی برای توصیف بهتر از ناحیه لب است، در مرحله دسته‌بندی می توان از دسته‌بندهای متفاوتی استفاده کرد. در این مقاله از درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان و k نزدیک‌ترین همسایه برای دسته‌بندی استفاده‌شده است.

ساختار مقاله به شرح زیر هست. بخش ۲، نحوه تشخیص ناحیه لب و استخراج نقاط کلیدی توضیح داده می‌شود. در بخش ۳، روش‌های استخراج ویژگی بررسی می‌شود. درنهایت نتایج به‌دست‌آمده مورد تحلیل قرار می‌گیرد.

۲- روش پیشنهادی

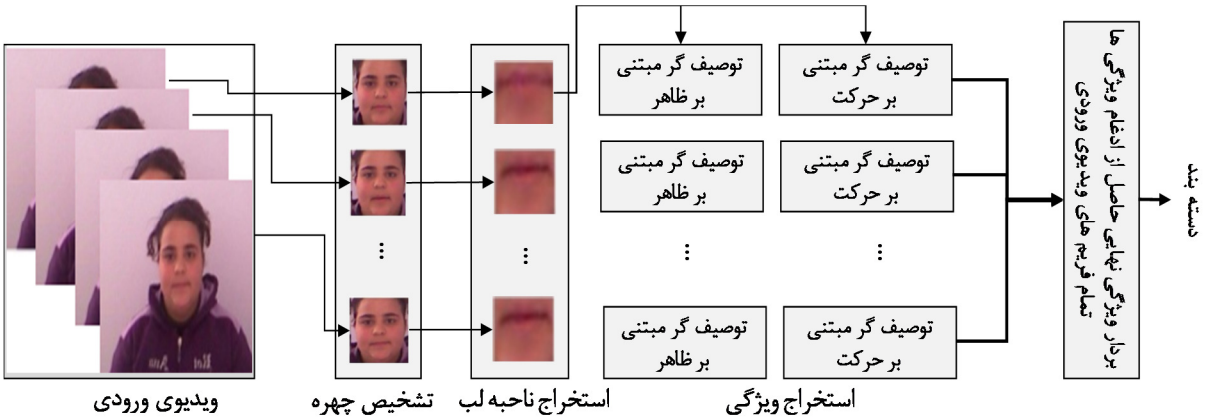
تغییرات فرم مرزهای لب دارای اطلاعات مهمی می‌باشند که می‌توان برای تشخیص گفتار از آن استفاده کرد. مرز لب در یک فریم حالت خاصی دارد که برای توصیف این حالت در این مقاله ویژگی‌های ظاهری بر پایه گرادیان تصویر پیشنهاد شده است. همچنین برای بهره‌مندی از ویژگی‌های مبتنی بر حرکت هیستوگرام گرادیان سه جهتِ (HOG3D)[۲۱] مورد استفاده قرار گرفت. برای توصیف حرکت نقاطِ کلیدی از برازش منحنی بر روی نقاط ناحیه لب و گرادیان در نقاط خاص از لب در طول زمان استفاده شد. به طور کلی، برای پیاده سازی روش ذکر شده مراحل زیر پیاده می‌شوند:

۱- تشخیص چهره در هر فریم از ویدیوی ورودی.
۲- استخراج ناحیه لب و ۳۸ نقطه کلیدی از این ناحیه.
۳- استخراج ویژگی‌های مبتنی بر ظاهر به ازای هر فریم.
۴ - استخراج منحنی توصیف کننده ویژگی های مکانی در طول فریم‌های ویدیو. در این بخش، نقطه مختصات یک نقطه کلید در تمام فریم‌ها به دست می‌آید و سپس یک منحنی درجه ۳ به این نقاط برازش می‌شود. ضرایب این منحنی به عنوان ویژگی در نظر گرفته می‌شود. این کار برای تمام نقاط کلیدی انجام می‌شود.
۵- ادغام ویژگی‌های لب در طول فریم‌های ویدیو.
۶- ارسال بردار ویژگی برای دسته‌بند و تشخیص گفتار ویدیو.

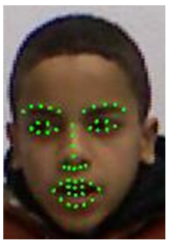
مراحل مختلف روش پیشنهادی در تصویر ۱ نشان داده‌شده است.

۲-۱- تشخیص ناحیه لب و استخراج نقاط کلیدی

برای تشخیص گفتار با استفاده از داده‌های بصری، ابتدا باید ناحیه دهان استخراج شود. در این مرحله ابتدا منطقه چهره با استفاده از الگوریتم ویولا و جونز [۲۲] شناسایی شد. روش ویولا و جونز به منظور محاسبه سریع ویژگی‌ها، تصویر تجمعی[۱۳] از روی تصاویر تولید می‌کند. ویژگی‌های شبه هار[۱۴] در چند مقیاس محاسبه و از کلاس بندهای ساده استفاده می‌کند. برای دستیابی به کلاس‌بندی سریع، تعداد بسیار زیاد از ویژگی‌ها حذف و بر بخش کوچک‌تر تمرکز می شود که این ویژگی‌ها با استفاده از الگوریتم یادگیری تقویت تطبیقی به دست می‌آید. سپس از ترکیب متوالی کلاس بندهای پیچیده به صورت آبشاری استفاده می‌کند. این کار باعث می‌شود آشکارساز بر روی مناطقی که احتمال وجود چهره بیشتر است تمرکز کرده و در نتیجه سرعت آشکارساز افزایش می‌یابد. برای تشخیص ناحیه لب‌ها روشی موسوم به چهرا[۱۵] [۲۳] استفاده شد. این روش ۴۹ نقطه کلید را روی چهره پیدا می‌کند. با استفاده از این نقاط می‌توان ویژگی های مورد نظر را برای نقاط کلیدی لب‌ها استخراج کرد (تصویر ۲).



تصویر ۱: مراحل پیاده‌سازی در روش پیشنهادی برای تشخیص گفتار با استفاده از پایگاه داده[۱۹]



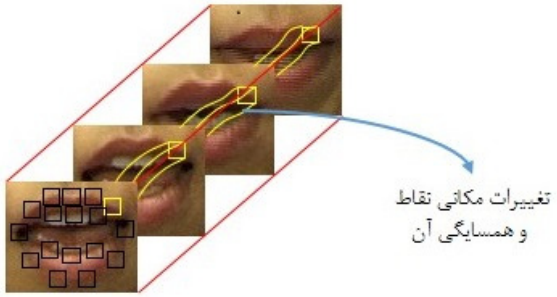
تصویر ۲: نقاطی که CHEHRA روی تصویر استخراج می‌کند[۲۰, ۲۳]

۲-۲- توصیف ناحیه لب

برای استخراج ویژگی‌های مناسب از روی تصاویر ناحیه دهان از دو روش که بر پایه گرادیان هستند و یک روش دنبال کردن نقاط کلیدی استفاده شد. روش اول جهت استخراج ویژگی‌های مکانی و روش دوم برای بهره‌مندی از ویژگی‌های زمانی و توالی است. گرادیان به جهت این‌که لبه‌های تصویر را استخراج می‌کند و برای لب‌خوانی تغییرات لبه‌ها، متمایزکننده‌ی گفتار بیان‌شده است، برای توصیف ناحیه لب روش مناسبی است.

۲-۳- گرادیان در نقاط کلیدی لب

گرادیان به دلیل استخراج لبه‌ها از تصویر می‌تواند در تشخیص گفتار مؤثر باشد. تغییرات مکانی نقاط ناحیه لب، ناشی از تغییرات لب‌ها در هنگام گفتار باعث تغییر در گرادیان‌ها می‌شود و درنتیجه الگوی خاصی را برای هر گفتار ایجاد می‌کند. برای استخراج ویژگی مناسب با استفاده از گرادیان، ابتدا گرادیان تصویر در دو جهت افقی و عمودی به دست آمد. در اطراف هر نقطه‌ی به‌دست‌آمده از روش CHEHRA که در اطراف لب‌ها قرار دارند، یک همسایگی ۹×۹ در نظر گرفته‌شده و در هر کدام از این نواحی، میانگین گرادیان افقی و عمودی محاسبه می‌گردد. تصویر ۳، هجده همسایگی در ناحیه لب و تغییرات مکانی این نواحی را نشان می‌دهد.



تصویر ۳: نواحی اطراف لب و تغییرات آن در طول فریم‌ها

به طور خلاصه گرادیان یک تصویر تغییر شدت رو شنایی یا رنگ یک تصویر را نشان می‌دهد. در هر تصویر، گرادیان یک نقطه، یک بردار دوبعدی است که از مشتق در دو جهت عمودی و افقی ایجاد می شود. متداول‌ترین روش برای محاسبه‌ی گرادیان تصویر اعمال آشکارسازهای لبه است. در این تحقیق از آشکار سازهای محاسبه گرادیان در جهت افقی و عمودی (تصویر ۴) استفاده شد، این فیلترها در تصویر اصلی کانوالو شده و دو ماتریس گرادیان افقی و عمودی ایجاد می‌کند.

$$\begin{bmatrix}1\\0\\-1\end{bmatrix} \qquad [1 \;\; 0 \;\; -1]$$

الف ب

تصویر ۴: فیلتر اعمالی برای محاسبه گرادیان عمودی و افقی

مراحل استخراج گرادیان را در تصویر ۵ مشاهده می‌کنید. پس از محاسبه میانگین گرادیان افقی و عمودی، از آنجایی‌که تعداد نقاط در نظر گرفته‌شده در ناحیه دهان ۱۸ نقطه بوده، از ۱۸ ناحیه‌ی موجود در محدوده لب‌ها، ۳۶ ویژگی برای هر فریم از ویدیوی ورودی استخراج می‌شود که بردار ویژگی حاصل به‌صورت زیر است:

$$FV_{gradient}=\left[\bar{G}^{v}_{p1},\bar{G}^{h}_{p1},\bar{G}^{v}_{p2},\bar{G}^{h}_{p2},\ldots,\bar{G}^{v}_{p18},\bar{G}^{h}_{p18}\right] \qquad (۱)$$

$\bar{G}$ میانگین گرادیان در محدوده مورد نظر است، v و h به ترتیب بیان‌کننده گرادیان در راستای عمودی و افقی می‌باشند و $p1$ تا $p18$ نشان‌دهنده نقاط به‌دست‌آمده در ناحیه لب هستند.

به طریق دیگری هم از این روش استخراج ویژگی استفاده شد، به این صورت که در ابتدای روش، تفاضل دو فریم متوالی به‌دست‌آمده و مراحل روش گرادیان در نقاط کلیدی بر روی تفاضل فریم‌ها طبق آنچه در بالا گفته شد، اعمال شد. با این کار ویژگی‌های زمانی که در تشخیص بصری گفتار مفید هستند نیز استخراج می‌شوند.

۲-۴- گرادیان زمانی

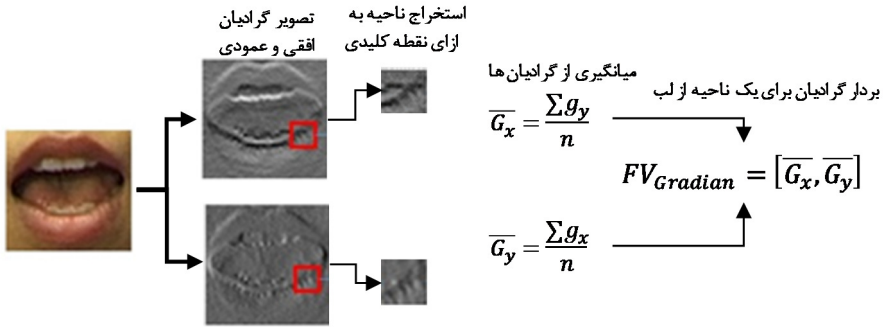
ویژگی‌های گرادیان تصویر هرچند قادراند فرم لب در یک فریم را نمایش‌دهند، اما عمده اطلاعات مفید برای لب‌خوانی در تغییرات نقاط کلیدی در طول زمان نهفته است. برای بهره‌گیری از این ویژگی مهم از گرادیان HOG3D [۲۱] استفاده شد که در ادامه این روش توضیح داده خواهد شد. در این روش، ابتدا گرادیان را در سه جهت محاسبه کرده و پس از آن انتگرال تصویر سه‌بعدی محاسبه می‌گردد (در تصویر ۶، C_i نشان‌دهنده سلول‌ها و b_i نشان‌دهنده زیر بلوک‌های آن است).

اگر برای یک ویدیو $v(x,y,t)$، گرادیان آن را در راستاهای x,y,t را با $v_{\partial x}v_{\partial y},v_{\partial t}$ نشان دهیم، انتگرال ویدیو برای $v_{\partial x}$ را می‌توان به صورت رابطه (۲) محاسبه کرد.

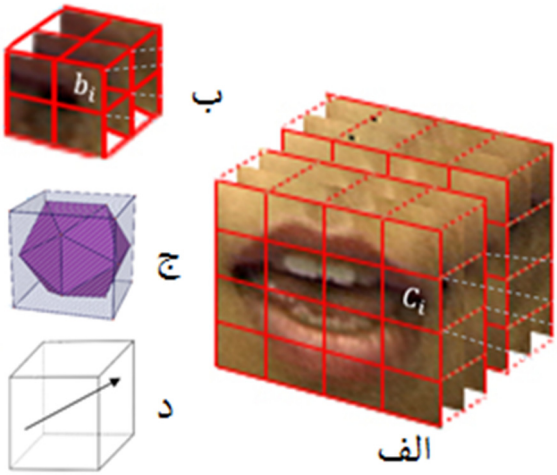
$$iv_{\partial t}=\sum_{x'\le x\,,\,y'\le y\,,\,t'\le t} v_{\partial t}(x',y',t') \qquad (۲)$$

حال برای یک زیر بلوک سه‌بعدی $b=(x,y,t,w,h,l)$ میانگین گرادیان باید محاسبه گردد. (x,y,t) مکان بلوک، w پهنا، h ارتفاع و l طول بلوک است و اگر میانگین گرادیان بلوک b در راستا $\bar{g}_b=(\bar{g}_{b\partial x},\bar{g}_{b\partial y},\bar{g}_{b\partial t})$ باشد، برای هرکدام از راستاها، میانگین گرادیان به صورت رابطه (۳) محاسبه می‌شود.

$$\bar{g}_{b\partial t}=\begin{bmatrix}iv_{\partial t}(x+w,y+h,t+l)-iv_{\partial t}(x,y+h,t+l)-\\ iv_{\partial t}(x+w,y,t+l)+iv_{\partial t}(x,y,t+l)\end{bmatrix} -\begin{bmatrix}iv_{\partial t}(x+w,y+h,t)-iv_{\partial t}(x,y+h,t)-\\ iv_{\partial t}(x+w,y,t)+iv_{\partial t}(x,y,t)\end{bmatrix} \qquad (۳)$$



تصویر ۵: مراحل استخراج ویژگی گرادیان در نقاط کلیدی لب از تصویر لب‌ها



تصویر ۶: مراحل روش HOG3D. الف) تقسیم لب به بلوک‌های سه‌بعدی. ب) تقسیم هرکدام از بلوک‌های سه‌بعدی به زیر بلوک‌ها و محاسبه هیستوگرام. ج) نگاشت بردار میانگین گرادیان هر زیر بلوک به یک‌چند وجهی و چندی کردن. د) بردار گرادیان حاصله از هر زیر بلوک برای توصیف حالت لب[۲۱]

در مرحله بعد مقدار میانگین گرادیان به‌دست‌آمده باید چندی بشود. برای این کار از اشکال سه بعدی چند وجهی استفاده شد. برای چندی‌کردن گرادیان سه بعدی، تصویر بردار $\bar{g}_b$ بر روی هر کدام از وجه‌های شکل سه بعدی چند وجهی محا سبه می شود. مقدار چندی شده روی وجه‌ها به‌صورت رابطه (۴) خواهد بود:

$$q_b=\left(q_{b1},q_{b2},\ldots,q_{bn}\right) \qquad (۴)$$

در این رابطه n تعداد وجه‌های ا ستفاده شده ا ست. برای محا سبه هیستوگرام گرادیان جهت‌ها، ابتدا سلول را به چند زیر بلوک به ابعاد $S\times S\times S$ تقسیم می‌کنیم. هر کدام از این زیر بلوک ها b_i نامیده شده‌اند و $\bar{g}_{b_i}$ میانگین گرادیان آن و q_{b_i} مقدار چندی شده هر بلوک است. اگر گرادیان سلول C_i را h_i بنامیم، h_i از مجموع مقادیر چندی شده q_{b_i} تمام زیر بلوک های آن محاسبه می‌شود. نهایتاً تمام هیستوگرام‌ها در یک بردار ویژگی ذخیره می‌شوند.

به طور خلاصه در این روش، ابتدا از تصاویر گرادیان سه‌بعدی و انتگرال سه بعدی گرفته می‌شود. سپس انتگرال حاصله را به سلول‌های سه‌بعدی تقسیم کرده و هر سلول را به زیر بلوک‌هایی تقسیم کرده و در هر زیر بلوک میانگین گرادیان در سه جهت محاسبه و بردار گرادیان حاصله را با تصویر کردن بردار بر روی صفحات یک چند وجهی، چندی کردیم. بردار گرادیان حا صل برای هر زیر بلوک محا سبه و هیستوگرام آن به عنوان ویژگی‌های سلول در نظر گرفته شد. بردار ویژگی حا صله به صورت رابطه (۵) خواهد بود:

$$FV_{hog3d}=\left[h_1,h_2,\ldots,h_M\right] \qquad (۵)$$

که h هیستوگرام هر سلول است و M تعداد سلول‌ها است که تأثیر این پارامتر در مسئله در بخش نتایج بیان می‌شود.

۲-۵- تخمین منحنی حرکت نقاط کلیدی لب

نقاط کلیدی ناحیه لب در طول زمان و در فریم‌های گفتار به‌صورت یک سیگنال در حال تغییر است. این سیگنال برای هر نقطه روی لب تغییرات مخصوص به خود را دارد و برای گفتار بیان شده با گفتار دیگر تفاوت دارد. از این تفاوت در تغییرات نقاط می‌توان ویژگی استخراج کرد. هر نقطه کلیدی از لب در طول زمان قابل رسم در یک نمودار ۳ بعدی مکان-زمان است. نحوه تغییرات این نقاط را می‌توان به کمک روش های رگرسیون تخمین زد. توابع چند جمله‌ای به وفور برای تخمین منحنی‌ها استفاده شده‌اند[۲۴]. در این پژوهش بر اساس مشاهدات از یک چندجمله‌ای درجه ۳ (برازش چندجمله‌ای) برای تخمین نحوه تغییرات نقاط کلیدی لب در طول زمان استفاده شد و برای این کار، بر هر یک از نقاط که از روش CHEHRA در ناحیه لب به‌دست‌آمده یک چند جمله‌ای درجه ۳ در طول فریم‌ها تطبیق داده می‌شود تا این چند جمله‌ای تغییرات نقطه را شبیه‌سازی کند. نهایتاً ضرایب این چندجمله‌ای تطبیق داده‌شده بر نقاط به عنوان ویژگی در نظر گرفته می‌شود. رابطه (۶) چند جمله‌ای درجه n را نشان می‌دهد.

$$f(x)=a_n x^n+a_{n-1}x^{n-1}+a_{n-2}\,x^{n-2}+\ldots+a_0 \qquad (۶)$$

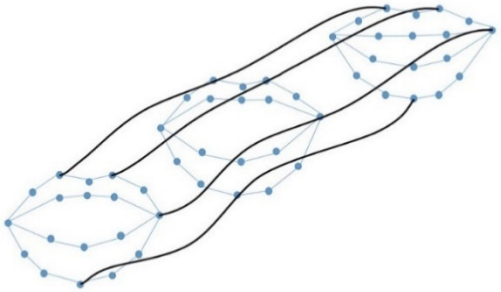
در معادله بالا مقدار n برابر ۳ بوده و چهار ضریب a_0,a_1,a_2,a_3 به عنوان ویژگی در نظر گرفته می‌شود.

به طور خلاصه در این روش، برای یک نقطه مانند A در ناحیه لب در تصویر ۷، تغییرات آن را با تطبیق یک چند جمله‌ای در جه ۳

مشابه‌سازی کرده و چهار ضریب این چند جمله‌ای به عنوان ویژگی در نظر گرفته شـــد، این کار برای هر ۱۸ نقطه ناحیه لب تکرار و برای هر گفتار بیان‌شـده ۷۲ ویژگی اســتخراج شـد که بردار ویژگی حاصــل به صورت رابطه (۷) است.

$$FV_{polygon}=\left[a_0^{p1},a_1^{p1},a_2^{p1},a_3^{p1},\ldots,a_1^{p18},a_2^{p18},a_3^{p18}\right] \quad (۷)$$

که $a_0.a_1.a_2.a_3$ ضـرایب حاصـله از برازش چند جمله‌ای بر روی یک نقطه است و $p1$ تا $p18$ نقاط ناحیه لب هستند.



تصویر ۷: نقاط ناحیه دهان و تغییرات آن در طول فریم‌ها استخراج شده از دادگان[۲۰]

ترکیب ویژگی‌های استخراج شده از روش گرادیان در نقاط کلیدی لب، گرادیان زمانی و تخمین منحنی حرکت نقاط کلیدی لب به عنوان ویژگی به دســته‌بندهای مختلف جهت بررســی و ارزیابی داده شــد و نتایج، دلایل و مشاهدات در بخش‌های بعد آورده شده است.

۲-۶- دسته‌بند

ویژگی‌های ذکرشده در بخش ۲ هر کدام به صورت جداگانه استخراج شده و به صورت جداگانه یا ادغام دو یا چند ویژگی به دسته‌بندها داده‌شدند. برای پیاده‌سازی از نرم‌افزار متلب استفاده شد و ابزار یادگیری ماشین مورد استفاده قرار گرفت. از آنجایی که هدف از این پژوهش ارائه روش استخراج ویژگی جدید بوده، از دسته‌بندهای موجود در ابزار یادگیری ماشین نرم‌افزار متلب استفاده شد و تغییری در پارامترهای دسته‌بندها داده نشد و از همان پارامترهای پیش‌فرض استفاده شد.

دسته‌بندهای مورد استفاده ماشین بردار پشتیبان (SVM)[۱۶] خطی، ماشین بردار پشتیبان با هسته چند جمله‌ای درجه سه (Cubic SVM)، K نزدیک‌ترین همســـایه (KNN)[۱۷]، درخت تصـــمیم (DT)[۱۸] و جنگل تصـادفی[۱۹] اسـت. در ادامه، هر کدام از این دسـته‌بندها، مزیت آن‌ها و دلیل استفاده از آن‌ها به طور مختصر توضیح داده خواهد شد.

SVM هر نقطه داده را به عنوان نقطه‌ای در فضای n بعدی ترسیم می‌کند که n تو سط تعدادی از ویژگی‌ها تعریف شده ا ست و ارزش هر یک از آن‌ها یک مقدار در مختصـات خاص اسـت. دسـته‌بندی با پیدا کردن ابر صفحه[۲۰] انجام می‌شود که دو کلاس را به بهترین شکل از هم تفکیک می‌کند. اگر ویژگی‌ها را نتوان در فضـــای n بعدی با یک ابر صـفحه خطی تفکیک کرد، می‌توان از هسـته[۲۱] برای نگاشـت داده‌ها به یک فضـای برداری دیگر که بتوان داده‌ها را با ابر صـفحه خطی تفکیک کرد، ا ستفاده نمود. در این مطالعه، ابتدا داده‌ها را با مدل خطی تفکیک شـدند و سـپس برای آزمون این موضـوع که SVM در فضـای برداری دیگر می‌توان تفکیک بهتری داشـته باشـد، تابع هسـته چند جمله‌ای درجه سه استفاده شد.

KNN یک الگوریتم ساده دیگر ا ست که تمام نمونه‌های موجود را به عنوان داده‌های آموز شی نگهداری می‌کند و نمونه جدید را بر ا ساس یک اندازه شباهت (مثلاً توابع فا صله) د سته‌بندی می‌کند. یک نمونه با رأی اکثریت همسایگان خود دسته‌بندی می‌شود و با اندازه‌گیری فاصله نمونه تا همسایگان، نزدیک‌ترین فاصله در میان K نزدیک‌ترین همسایه کلاس نمونه را تعیین می‌کند. از آنجایی که در لب‌خوانی تغییرات در رو شنایی و حالت‌های حرکات لب‌ها در هنگام گفتار در انسان‌ها نسبت به هم متفاوت اسـت[۱]، از این خصـوصـیت رأی اکثریت KNN برای بررســی تأثیر رأی اکثریت اســتفاده شــد که آیا این چالش را برطرف می‌کند. در این مطالعه K = 1 انتخاب شده است.

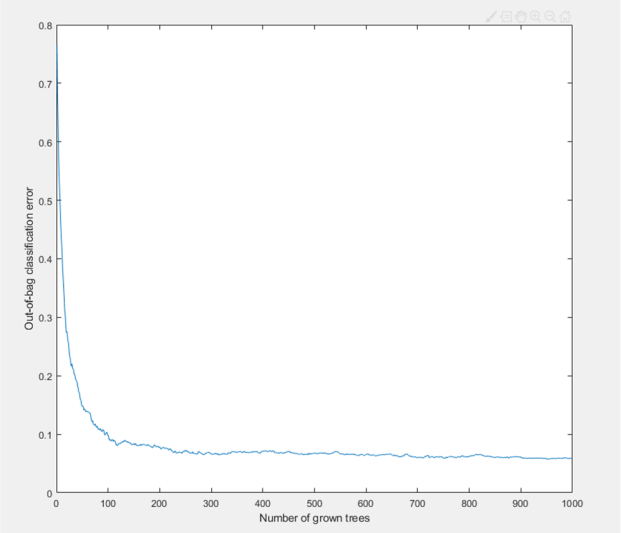
DT مدلی ا ست م شابه فلوچارت، ساختاری درخت مانند را جهت اخذ ت صمیم و تعیین کلاس و د سته یک داده خاص به ما ارائه می‌کند. همان‌طور که از نام آن مشـــخص اســـت، این درخت از تعدادی گره و شاخه تشکیل‌شده است به‌گونه‌ای که‌برگ‌ها کلاس‌ها یا دسته‌بندی‌ها را ن شان می‌دهند و گره‌های میانی هم برای ت صمیم‌گیری با توجه به یک یا چند صـــفت خاص به کار می‌روند. توجه مهم در مورد درخت‌های ت صمیم گیر این ا ست که آن‌ها در برابر اختلال و نویز قوی ه ستند. در مورد تشخیص گفتار، جایی که منطقه موردعلاقه در تصویر شامل دهان معمولاً یک منطقه کوچک اســـت، نویز می‌تواند نقش مهمی در نتایج دســـته‌بندی بازی کند[۱]؛ بنابراین، اثربخشـــی این دســـته‌بند برای تشخیص بصری گفتار در این مطالعه مورد بررسی قرار گرفت.

جنگل تصـادفی یک گروه از درخت‌های تصـمیم اسـت. اسـتراتژی ساخت این گروه در افزایش تنوع در میان درختان متمرکز شده ا ست. درخت‌های تصـمیم‌گیر بسـیار ناپایدار هسـتند به این صـورت که یک تغییر کوچک در مجموعه داده، تغییرات زیادی در مدل توســعه یافته ایجاد می‌کند[۱۱]. به این دلیل از جنگل تصـــادفی در این پژوهش استفاده شد تا ناپایداری درخت تصمیم را نداشته باشد و با رأی اکثریت درخ تان پیش‌بینی صـــورت پذیرد. برای هر نود این روش یک زیر مجموعه کوچک ت صادفی از ویژگی‌ها انتخاب می شود و تنها از این زیر مجموعه برای بهترین تقسـیم‌بندی جسـتجو می‌شـود. تعداد درخت در این بررسـی ۱۰۰۰ عدد در نظر گرفته شـد که بر اسـاس میزان خطای جنگل به ازای تعداد درخت‌ها این عدد انتخاب شد (تصویر ۸).

۳- نتایج

در این ارز یابی از داده‌های MIRACL-VC1 اســـتفاده شـــد که یک مجموعه داده لب‌خوانی اسـت که دارای ۱۰ کلمه (۱۵ نفر × ۱۰ حرف

× ۱۰ تکرار) و ۱۰ عبارت (۱۵ نفر × ۱۰ عبارات × ۱۰ تکرار) است.



تصویر ۸: نسبت خطا به رشد تعداد درختان، محور افقی تعداد درختان و محور عمودی میزان خطا است.

پایگاه داده با استفاده از ابزار کینکت که دارای یک دوربین معمولی و یک دوربین برای ثبت عمق تصویر است تصویربرداری شد. در این پایگاه داده ۳ هزار نمونه وجود دارد که ۱۵۰۰ نمونه بر روی ۱۰ کلمه و ۱۵۰۰ نمونه بر روی ۱۰ عبارت است. هر کلمه توسط ۱۵ نفر، ۱۰ بار تکرار و تصویربرداری شده است. این کار برای عبارات هم انجام شده است. این مجموعه شامل کلمات مانند navigation, connection و غیره و عبارات روزمره مانند Nice to meet you, I love this game و غیره است. در این مطالعه تنها تصاویر رنگی که توسط Kinect گرفته‌شده‌اند مورد بررسی قرار گرفتند. ایجادکننده مجموعه داده، تصاویر عمق را عمدتاً برای استخراج ناحیه لب در تصویر استفاده می‌کند. بااین‌حال مشاهدات ما نشان داد که تنها با استفاده از تصاویر رنگی و الگوریتم Viola-Jones برای هدف این مطالعه کافی است. چند آزمون برای ارزیابی انجام شد:

- آزمون مستقل از گوینده (SI)
- آزمون وابسته به گوینده (SD)

در آزمون SI، استراتژی کنار گذاشتن یک گوینده استفاده شد که اطلاعاتی از یک گوینده به‌عنوان داده‌های آزمون استفاده شد و داده‌های گویندگان باقیمانده برای مرحله آموزش مورد استفاده قرار گرفت. همین روش برای هر گوینده در مجموعه داده تکرار شد. در روش آموزش SI داده‌های آموزش و آزمون به این صورت جدا می‌شود که تمام ویدیوهای مربوط به یک گوینده برای آزمون و ویدیو مابقی گویندگان برای آموزش در نظر گرفته می‌شوند و این کار به تعداد گویندگان تکرار می‌شود یعنی یک بار که ویدیوهای یک گوینده به عنوان آزمون در نظر گفته شد، در بار بعد ویدیوی گوینده دیگر برای آزمون قرار داده می‌شود. نهایتاً به تعداد گویندگان آموزش و آزمون صورت می‌گیرد و نتایج از میانگین‌گیری نتایج هر بار آموزش و آزمون به دست می‌آید.

در آزمون SD، آموزش و داده‌های آزمون از همان گوینده به دست آمد. برای هر یک از گویندگان در مجموعه داده‌ها، اعتبار سنجی متقابل از ویدئوهای گوینده انجام شد، یعنی دو ویدیو برای آزمون استفاده شد و بقیه برای آموزش استفاده شد. روش آموزش SD با استفاده از k-fold انجام شده و تعداد k عدد ۱۰ در نظر گرفته شد. این آزمون‌ها در [۱]، [۱۱] و [۲۰] نیز مورد استفاده قرار گرفتند.

نتایج این مقاله بر اساس معیار دقت[۲۲]، صحت[۲۳]، بازخوانی[۲۴] و معیار F[۲۵] است که در رابطه (۸) تا (۱۱) مشاهده می‌کنید:

$$average\_Accuracy = \frac{\sum_{i=1}^{L} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}}{L} \quad (۸)$$

$$\Pr ecision = \frac{\sum_{i=1}^{L} \frac{tp_i}{tp_i + fp_i}}{L} \quad (۹)$$

$$\Re call = \frac{\sum_{i=1}^{L} \frac{tp_i}{tp_i + fn_i}}{L} \quad (۱۰)$$

$$F1\_score = \frac{2 \times (precision \times recall)}{precision + recall} \quad (۱۱)$$

در این چهار رابطه، L تعداد کلاس‌ها، tp_i تعداد مشاهدات به درستی تشخیص داده شده برای کلاس C_i، tn_i تعداد مشاهدات به درستی تشخیص داده شده که متعلق به کلاس C_i نیستند، fp_i تعداد مشاهداتی که به نادرستی تشخیص داده شده که متعلق به کلاس C_i هستند، fn_i تعداد مشاهداتی که تشخیص داده نشده که متعلق به کلاس C_i هستند، است.

برای مقایسه با روش‌های پیشین، تعدادی از روش‌ها را پیاده‌سازی کردیم که برخی از جزییات این روش‌ها در ادامه بیان خواهد شد. برای روش HOG سایز سلول ۸×۸ و اندازه بلوک ۱۶×۱۶ و گام لغزش بلوک ۸×۸ گزارش شده است که بر روی تمام نواحی دهان اعمال شده است. برای روش lbp اندازه پنجره ۳×۳ بوده و بر روی تمام ناحیه دهان اعمال شده است. ویژگی‌های ذکرشده در مقالات پیشین، به کمک کتابخانه‌های موجود در نرم‌افزار متلب پیاده‌سازی نمودیم و روش استخراج ویژگی آن‌ها به جای روش استخراج ویژگی پیشنهادی جایگزین کرده و باقی مراحل مطابق با روش ارائه شده در این پژوهش انجام شده است.

اگر بخواهیم از نظر طول بردار ویژگی مقایسه کنیم، طول بردار ویژگی حاصله برای هر ویدیو از روش استخراج ویژگی گرادیان در نقاط کلیدی و روش گرادیان تفاضل تصاویر در نقاط کلیدی هر کدام ۷۲۰ ویژگی، برای روش گرادیان زمانی ۳۸۸۸ ویژگی و برای روش برازش منحنی ۷۲ ویژگی است. به طور کلی طول بردار ویژگی حاصل از ادغام روش‌های استخراج ویژگی ۵۴۰۰ ویژگی است که در مقایسه با مقالاتی مانند [۱] و [۲۰] که برای هر ویدیو ۱۹۴۴۰ ویژگی استخراج می‌کنند، کوچک است.

اگر بخواهیم تعداد تکرار روش (پیچیدگی زمانی) را محاسبه کنیم، با فرض اینکه n تعداد نمونه‌ها و d تعداد ویژگی‌ها است، برای استخراج ویژگی به تعداد نمونه‌ها یک حلقه وجود دارد O(n) و برای دسته‌بندی هر کدام از دسته‌بندها پیچیدگی زمانی خود را دارند و به عنوان نمونه، جنگل تصادفی دارای پیچیدگی زمانی $O(ntree \times mtry \times n \times \log(n))$ است که در این جا ntree تعداد درختان و mtry تعداد ویژگی‌هایی است که در هر گره درخت نمونه برداری می‌شود.

۳-۱- نتایج آزمایش بر روی کلمات

جدول ۱ نتایج ارزیابی تشخیص بصری گفتار با استفاده از ویژگی‌ها و دسته‌بندهای ذکرشده برای کلمات نشان می‌دهد. این جدول نتایج آزمون با توجه به آزمون مستقل از گوینده را نشان می‌دهد.

طبق این نتایج، بیشترین دقت با ترکیب همه‌ی روش های استخراج ویژگی بوده و همچنین دقت روش استخراج ویژگی گرادیان زمانی در نقاط کلیدی لب و HOG3D، با استفاده از روش دسته‌بندی جنگل تصادفی حدود ۶۹٪ به دست آمد. به صورت کلی جنگل تصادفی (RF) بهترین دسته‌بند بوده و به صورت میانگین حدود ۶۲/۹٪ دقت در تشخیص و دسته‌بندی داشته است. جنگل تصادفی به دلیل این که از چندین درخت تصمیم در خود بهره می‌برد و در آن هر کدام از درختان تعدادی از نمونه ها را به عنوان داده های آزمون در یافت می‌کنند؛ درنتیجه جنگل تصادفی نسبت به درخت تصمیم (DT) که به طور میانگین حدود ۳۰٪ دقت داشته، توانسته به‌دقت بالایی برسد. ماشین بردار پشتیبان بعد از جنگل تصادفی به دقت مطلوبی رسیده است. ترکیب روش استخراج ویژگی تطبیق چند جمله‌ای با روش گرادیان در نقاط کلیدی لب توانسته دقت روش گرادیان در نقاط کلیدی لب را بین ۲ تا ۶ درصد افزایش دهد که این نشان دهنده تأثیرگذاری روش‌های استخراج ویژگی مبتنی بر حرکت در بهبود تشخیص است.

به‌صورت کلی، ترکیب روش‌های استخراج ویژگی ارائه‌شده، توانسته به‌طور میانگین به‌دقت ۴۷٪ برسد و پس از آن ترکیب روش‌های گرادیان در نقاط کلیدی لب و گرادیان زمانی در نقاط کلیدی لب، به دقت حدود ۴۶٪ رسید. استفاده از ترکیب همه روش‌ها به دلیل بهره‌مندی و استفاده از ویژگی‌های زمانی (ویژگی‌های مبتنی بر حرکت) و مکانی (ویژگی های مبتنی بر ظاهر) توانسته دقت بالایی را در تشخیص کلمات داشته باشد.

روش پیشنهادی در دسته‌بند جنگل تصادفی، نسبت به روش استخراج ویژگی DCT که بیشترین دقت را در بین روش‌های قبلی مورد استفاده در لب‌خوانی دارد، حدوداً ۱۰٪ دقت بیشتری دارد. ترکیب دو روش استخراج ویژگی MBH و HOG مورد استفاده در مقاله [۱] با وجود بهره‌مندی از ویژگی‌های ظاهری و ویژگی‌های حرکتی نتوانسته بر روش پیشنهادی برتری داشته باشد و حدود ۱۷٪ دقت کمتری نسبت به روش پیشنهادی دارد.

همچنین ترکیب روش‌های پیشنهادی حدوداً ۷۸ درصد معیار F آن است که در مقایسه با روش DCT، ۱۲ درصد بهبود داشته است.

جدول ۲ نتایج را برای ارزیابی تشخیص گفتار بصری با استفاده از ویژگی‌ها و دسته‌بندهای ذکرشده برای کلمات نشان می‌دهد. این جدول نتایج آزمون با توجه به آزمون وابسته به گوینده را نشان می‌دهد و چون در داده‌های آموزش دسته‌بند، نمونه‌ای از کلمه‌ی بیان‌شده توسط تمام کاربرها شرکت دارد، صحت بالایی نسبت به روش آزمون مستقل از گوینده دارد.

به‌صورت کلی روش طبقه بند SVM با هسته چند جمله‌ای به طور میانگین به دقت حدود ۸۵٪ رسیده و RF حدوداً دقت ۸۲٪ دارد. روش استخراج ویژگی گرادیان در نقاط کلیدی لب بالاترین دقت را به‌طور میانگین در تشخیص داشته و همچنین ترکیب روش گرادیان در نقاط کلیدی لب با روش تطبیق چند جمله‌ای به همان اندازه استفاده از گرادیان در نقاط کلیدی لب دقت در تشخیص داشته است. بعد از آن ترکیب گرادیان در نقاط کلیدی لب و گرادیان زمانی در نقاط کلیدی لب، توانسته به خوبی دقت در تشخیص را افزایش دهد.

در بین روش‌های پیشین، DCT با اینکه در آزمون مستقل از گوینده دقت بالایی داشته ولی در آزمون وابسته به گوینده، دقت آن کم شده است.

ترکیب روش های HOG و MBH بالاترین دقت را در روش های پیشین داشته ولی با این حال نتوانسته از روش‌های پیشنهادی سبقت بگیرد.

جدول ۱: نتایج تشخیص گفتار بر روی کلمات و با آزمون مستقل از گوینده

ویژگی‌ها	معیارها	DT	Linear SVM	Cubic SVM	KNN	RF
$FV_{gradient}^{spatial}$	Precision	۳۳/۸۷	**۶۲/۱۶**	۵۹/۱۸	۳۷/۵۶	۵۷/۷۴
	Recall	۶۷/۶۹	۸۴/۲۶	۸۰/۷۰	۶۳/۹۰	۸۰/۶۷
	F1-score	۴۳/۳۴	۶۹/۰۷	۶۵/۵۲	۴۵/۱۵	۶۵/۱۰
$FV_{gradient}^{temporal}$	Precision	۲۷/۷۸	۵۹/۴۵	۶۲/۰۰	۴۷/۷۰	۶۰/۶۷
	Recall	۶۹/۰۰	۸۸/۸۵	۸۹/۳۶	۸۴/۱۲	۹۰/۲۱
	F1-score	۳۸/۳۷	۶۹/۲۳	۷۱/۱۸	۵۹/۱۹	۷۱/۰۷
FV_{hog3d}	Precision	۳۰/۹۴	۵۷/۱۱	۶۰/۷۴	۴۶/۶۰	۶۵/۷۴
	Recall	۶۹/۱۰	۸۴/۳۱	۸۶/۱۱	۷۹/۸۸	۸۹/۰۳
	F1-score	۴۱/۳۶	۶۶/۱۸	۶۹/۴۲	۵۶/۹۸	۷۳/۹۱
$FV_{polygon}$	Precision	۲۴/۹۴	۸/۰۰	۳۳/۵۷	۲۲/۹۷	۳۶/۳۳
	Recall					

	F1-score	۶۲/۵۴	۲۱/۶۲	۶۲/۵۰	۵۸/۵۱	۷۰/۰۹
		۳۴/۵۰	۱۰/۲۳	۴۱/۶۱	۳۱/۷۱	۴۵/۶۱
$FV_{gradient}^{spatial} + FV_{polygon}$	Precision	۳۶/۹۹	۳۸/۵۷	۳۸/۵۷	۳۸/۰۴	۶۳/۵۶
	Recall	۷۱/۴۰	۵۶/۹۲	۵۶/۹۲	۶۵/۱۷	۸۶/۱۷
	F1-score	۴۷/۱۳	۴۳/۲۰	۴۳/۲۰	۴۵/۸۷	۷۱/۳۳
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal}$	Precision	۳۳/۸۶	۵۲/۹۱	۵۲/۹۲	۴۷/۶۰	۶۴/۴۰
	Recall	۶۹/۱۵	۸۳/۲۲	۸۳/۲۸	۸۴/۰۸	۸۸/۷۹
	F1-score	۴۳/۹۰	۶۲/۶۵	۶۲/۶۹	۵۹/۰۹	۷۲/۹۳
$FV_{gradient}^{spatial} + FV_{hog3d}$	Precision	۳۴/۶۷	۵۳/۳۱	۵۳/۳۵	۴۷/۶۰	۶۴/۳۵
	Recall	۷۰/۷۷	۸۰/۴۰	۸۰/۴۱	۷۹/۷۸	۸۶/۵۴
	F1-score	۴۴/۸۹	۶۱/۹۵	۶۱/۹۸	۵۷/۸۰	۷۱/۸۳
$FV_{gradient}^{temporal} + FV_{hog3d}$	Precision	۲۴/۹۴	۵۶/۶۲	۵۶/۶۰	۵۲/۴۳	۶۹/۴۸
	Recall	۶۶/۷۷	۸۵/۹۶	۸۵/۸۳	۸۵/۵۳	**۹۲/۶۱**
	F1-score	۳۸/۳۱	۶۶/۲۲	۶۶/۱۱	۶۳/۵۲	۷۷/۸۸
$FV_{hog3d} + FV_{polygon}$	Precision	۳۳/۲۵	۴۶/۲۰	۴۶/۲۰	۴۶/۴۵	۶۸/۰۴
	Recall	۷۰/۴۶	۷۶/۷۰	۷۶/۷۰	۷۹/۸۵	۹۰/۶۸
	F1-score	۴۳/۸۱	۵۵/۵۸	۵۵/۵۸	۵۶/۸۸	۷۵/۸۶
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{hog3d}$	Precision	۳۲/۸۲	۶۰/۳۷	۶۰/۳۳	۵۴/۴۸	۶۸/۱۰
	Recall	۶۹/۴۷	۸۶/۴۹	۸۶/۴۶	۸۵/۵۹	۹۰/۴۲
	F1-score	۴۳/۲۳	۶۹/۰۶	۶۹/۰۲	۶۴/۸۸	۷۵/۸۶
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{polygon}$	Precision	۳۶/۸۸	۵۹/۱۵	۵۹/۱۵	۵۲/۱۴	۶۷/۲۹
	Recall	۷۳/۰۸	۸۳/۵۹	۸۳/۵۹	۸۵/۹۸	۹۰/۴۲
	F1-score	۴۷/۸۴	۶۶/۹۲	۶۶/۹۲	۶۲/۷۹	۷۵/۳۷
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{polygon} + FV_{hog3d}$	Precision	۳۴/۴۵	۶۰/۴۵	۶۰/۴۵	۵۴/۶۱	**۷۱/۲۹**
	Recall	۷۰/۲۹	۸۶/۵۲	۸۶/۵۲	۸۵/۶۹	۹۲/۱۴
	F1-score	۴۴/۷۴	۶۹/۱۴	۶۹/۱۴	۶۵/۰۴	**۷۸/۸۸**
DCT [12]	Precision	۲۸/۷۶	۶۲/۰۴	۵۶/۹۶	۳۹/۷۱	۵۹/۲۶
	Recall	۵۹/۸۴	۸۳/۷۷	۸۱/۰۰	۶۴/۳۵	۸۱/۶۷
	F1-score	۳۷/۲۰	۶۸/۹۰	۶۴/۷۶	۴۷/۰۳	۶۶/۵۳
HOG [1]	Precision	۲۲/۲۵	۵۵/۸۵	۵۵/۷۴	۳۵/۲۸	۵۲/۹۴
	Recall	۵۶/۳۴	۷۹/۶۵	۷۹/۹۸	۶۲/۵۲	۷۷/۲۶
	F1-score	۳۰/۹۱	۶۳/۳۸	۶۳/۲۵	۴۲/۹۸	۶۰/۵۴
LBP [11]	Precision	۲۳/۸۳	۴۹/۵۷	۵۰/۸۸	۳۱/۸۴	۵۱/۹۷
	Recall	۶۲/۱۷	۷۵/۰۴	۷۴/۲۶	۵۹/۱۲	۷۶/۹۱
	F1-score	۳۳/۶۳	۵۷/۳۴	۵۷/۸۹	۳۹/۴۰	۵۹/۷۹
MBH [1]	Precision	۱۲/۶۳	۳۸/۰۰	۳۹/۷۱	۱۸/۴۹	۳۲/۸۰
	Recall	۴۴/۴۷	۷۲/۲۸	۷۳/۲۴	۵۱/۱۷	۶۹/۴۴
	F1-score	۱۹/۳۳	۴۸/۰۴	۴۹/۷۹	۲۶/۴۶	۴۲/۹۵
MBH + HOG [1]	Precision	۱۸/۹۰	۵۹/۸۸	۶۰/۹۸	۳۱/۴۳	۵۴/۱۹
	Recall	۵۰/۸۵	۸۵/۵۰	۸۵/۵۴	۶۷/۲۱	۷۸/۱۵
	F1-score	۲۶/۷۵	۶۷/۷۲	۶۸/۸۱	۴۱/۰۳	۶۱/۷۲

جدول ۲: نتایج تشخیص گفتار بر روی کلمات و با آزمون وابسته به گوینده و معیار صحت

ویژگی‌ها	DT	Linear SVM	Cubic SVM	KNN	RF
$FV_{gradient}^{spatial}$	۵۲/۸۰	۸۵/۵۰	**۹۳/۱۰**	۹۴/۱۰	۹۲/۴۰
$FV_{gradient}^{temporal}$	۳۷/۷۰	۷۴/۸۰	۸۱/۹۰	۸۴/۴۰	۷۷/۷۰
FV_{hog3d}	۴۲/۸۰	۸۵/۲۰	۸۶/۱۰	۷۸/۸۰	۶۹/۹۰
$FV_{polygon}$	۳۳/۲۰	۴۸/۷۰	۴۸/۲۰	۳۸/۳۰	۴۶/۷۰

$FV_{gradient}^{spatial}+FV_{polygon}$	۵۵/۴۰	۸۷/۲۰	۹۲/۴۰	۹۳/۷۰	۹۰/۵۰
$FV_{gradient}^{spatial}+FV_{gradient}^{temporal}$	۵۲/۶۰	۸۵/۵۰	۹۱/۵۰	۹۳/۵۰	۹۰/۳۰
$FV_{gradient}^{spatial}+FV_{hog3d}$	۵۱/۵۰	۸۷/۸۰	۸۹/۴۰	۸۳/۱۰	۹۰/۹۰
$FV_{gradient}^{temporal}+FV_{hog3d}$	۴۰/۱۰	۸۶/۹۰	۸۹/۲۰	۸۳/۰۰	۸۶/۹۰
$FV_{hog3d}+FV_{polygon}$	۴۵/۸۰	۸۴/۹۰	۸۷/۱۰	۷۴/۷۰	۸۴/۱۰
$FV_{gradient}^{spatial}+FV_{gradient}^{temporal}+FV_{hog3d}$	۵۱/۳۰	۸۹/۳۰	۹۱/۵۰	۸۷/۷۰	۹۰/۹۰
$FV_{gradient}^{spatial}+FV_{gradient}^{temporal}+FV_{polygon}$	۵۲/۷۰	۸۶/۷۰	۹۱/۵۰	۹۳/۱۰	۹۰/۳۰
$FV_{gradient}^{spatial}+FV_{gradient}^{temporal}+FV_{polygon}+FV_{hog3d}$	۵۰/۰۰	۸۹/۷۰	۹۱/۴۰	۸۷/۵۰	۹۰/۷۰
DCT [12]	۴۱/۴۰	۵۴/۶۰	۵۶/۴۰	۷۵/۷۰	-
HOG [1]	۳۶/۵۰	۸۵/۹۰	۹۲/۶۰	۸۷/۱۰	-
LBP [11]	۲۰/۷۰	۵۳/۴۰	۶۷/۵۰	۶۸/۶۰	-
MBH [1]	۳۱/۲۰	۸۳/۲۰	۸۹/۸۰	۸۸/۲۰	-
MBH + HOG [1]	۳۳/۴۰	۸۷/۷۰	۹۲/۴۰	۸۸/۸۰	-

۳-۲- نتایج آزمایش بر روی جملات

در روش پیشنهادی برای تشخیص جملات، ویژگی‌های هر فریم به صورت جداگانه استخراج می‌شود. سپس بردارهای ویژگی در کنار یکدیگر قرار می‌گیرند و به کمک روش درون‌یابی طول بردارهای ویژگی نرمال می‌شود. جدول ۳ نتایج مربوط به تشخیص جملات و با آزمون مستقل از گوینده (SI) را نشان می‌دهد. بیشترین دقت حدود ۷۶٪ در تشخیص با دسته‌بند جنگل تصادفی و با ترکیب روش‌های استخراج ویژگی گرادیان در نقاط کلیدی لب، گرادیان زمانی در نقاط کلیدی لب و HOG3D حاصل شد. RF به مانند آزمایش بر روی کلمات بیشترین دقت را در میان دسته‌بندها داشت که دلیل آن هم به خاطر استفاده از چندین درخت در خود و رأی‌گیری برای تعیین کلاس است.

ترکیب روش‌های استخراج ویژگی نیز به مانند آزمایش بر روی کلمات نسبت به سایر روش‌ها از دقت مطلوبی برخوردار است. در این آزمایش به دلیل این که تعداد فریم‌های جملات بیشتر از کلمات هستند، مشاهده می‌شود که نتایج آزمون بر روی جملات حدود ۸٪ بهتر از نتایج آزمون بر روی کلمات است.

بیشترین مقدار معیار F را ترکیب روش‌های گرادیان در نقاط کلیدی، گرادیان زمانی در نقاط کلیدی و HOG3D داشته و نسبت به روش‌های پیشین حدوداً ۱۰٪ بهتر است. همچنین ترکیب این روش‌ها بیشترین میزان فراخوانی را دارد.

از نظر دقت روش پیشنهادی بهتر از روش‌های پیشین خصوصاً DCT داشته و بین ۲ تا ۱۰ درصد بهتر عمل کرده است.

جدول ۴ نتایج تشخیص جملات و تحت آزمون وابسته به گوینده (SD) را نشان می‌دهد. SVM چند جمله‌ای و روش استخراج ویژگی گرادیان در نقاط کلیدی لب با صحت حدوداً ۹۵٪ بیشترین مقدار معیار صحت را دارد. روش استخراج ویژگی گرادیان در نقاط کلیدی لب به تنهایی صحت بالایی در تشخیص داشته و بعد از آن ترکیب گرادیان در نقاط کلیدی لب و تطبیق چند جمله‌ای، توانسته به خوبی معیار صحت در تشخیص را افزایش دهد.

ترکیب دو روش HOG و MBH بیشترین صحت را در بین روش‌های پیشین داشته است و با اینکه از ویژگی‌های حرکتی بهره می برد نسبت به روش پیشنهادی صحت کمتری دارد. در بعضی از موارد صحت روش پیشنهادی بیشتر نیز بوده و در بیشتر موارد رقابت پایا پایی داشته است. به‌عنوان نمونه ترکیب روش‌های گرادیان در نقاط کلیدی و تطبیق چند جمله‌ای با دسته‌بند SVM و هسته چند جمله‌ای برتری نسبی نسبت به ترکیب دو روش HOG و MBH دارد.

جدول ۳: نتایج تشخیص گفتار بر روی جملات و با آزمون مستقل از گوینده

ویژگی‌ها	معیارها	DT	Linear SVM	Cubic SVM	KNN	RF
$FV_{gradient}^{spatial}$	Precision	۳۸/۸۵	۷۱/۱۰	۷۱/۵۵	۵۵/۷۵	۷۱/۸۴
	Recall	۶۹/۱۴	۸۹/۴۷	۸۹/۶۸	۸۰/۴۹	۸۸/۸۴
	F1-score	۴۷/۸۹	۷۷/۴۱	۷۷/۹۰	۶۳/۳۵	۷۷/۹۳
$FV_{gradient}^{temporal}$	Precision	۲۷/۸۳	۶۶/۴۰	۷۱/۲۴	۵۸/۱۷	۶۸/۰۴
	Recall					

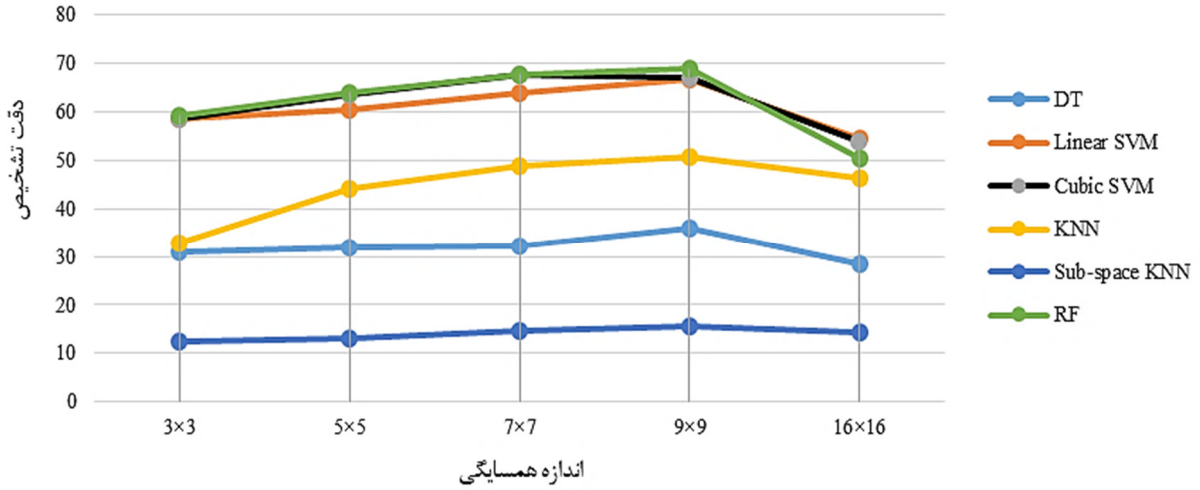
	F1-score	۶۸/۸۰	۹۰/۸۵	۹۲/۱۸	۸۸/۹۳	۹۰/۶۰
		۳۸/۵۱	۷۵/۲۱	۷۸/۸۶	۶۸/۸۱	۷۶/۵۰
FV_{hog3d}	Precision	۲۸/۰۳	۶۲/۶۴	۶۶/۷۴	۴۸/۵۹	۶۷/۶۳
	Recall	۶۵/۶۷	۸۷/۲۷	۹۰/۳۹	۸۳/۰۸	۹۰/۳۶
	F1-score	۳۸/۱۰	۷۰/۹۳	۷۵/۱۵	۵۹/۷۱	۷۵/۹۶
$FV_{polygon}$	Precision	۲۵/۰۸	۱۱/۵۴	۳۹/۵۳	۲۶/۱۹	۳۲/۰۴
	Recall	۶۰/۲۹	۲۷/۴۱	۶۸/۸۹	۶۲/۶۲	۶۴/۴۰
	F1-score	۳۴/۱۸	۱۴/۹۸	۴۸/۰۵	۳۵/۵۸	۴۱/۱۲
$FV_{gradient}^{spatial} + FV_{polygon}$	Precision	۳۸/۲۲	۵۳/۰۲	۵۳/۰۲	۵۶/۲۸	۶۷/۱۶
	Recall	۶۷/۵۹	۷۴/۰۲	۷۴/۰۲	۸۱/۱۲	۸۷/۳۳
	F1-score	۴۶/۶۴	۵۹/۲۲	۵۹/۲۲	۶۳/۹۱	۷۴/۳۲
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal}$	Precision	۳۹/۱۱	۶۸/۵۱	۶۸/۵۱	۶۳/۳۸	۷۴/۷۵
	Recall	۷۱/۸۸	۸۹/۰۲	۸۹/۰۲	۹۰/۴۳	۹۰/۵۷
	F1-score	۴۸/۶۳	۷۵/۴۸	۷۵/۴۸	۷۲/۷۱	۸۰/۵۷
$FV_{gradient}^{spatial} + FV_{hog3d}$	Precision	۳۸/۲۳	۶۲/۲۱	۶۲/۲۱	۵۴/۸۷	۷۴/۱۲
	Recall	۷۱/۳۳	۸۶/۹۴	۸۶/۹۴	۸۳/۹۳	۹۲/۰۷
	F1-score	۴۷/۹۲	۷۰/۶۶	۷۰/۶۶	۶۴/۳۷	۸۰/۸۵
$FV_{gradient}^{temporal} + FV_{hog3d}$	Precision	۳۱/۹۷	۶۴/۲۶	۶۴/۳۵	۶۲/۰۱	۷۲/۳۴
	Recall	۶۸/۷۷	۸۹/۳۹	۸۹/۴۴	۸۹/۵۷	۹۲/۵۰
	F1-score	۴۱/۹۵	۷۳/۱۰	۷۳/۱۹	۷۱/۸۶	۷۹/۹۰
$FV_{hog3d} + FV_{polygon}$	Precision	۲۹/۵۳	۵۳/۶۶	۵۳/۶۶	۴۸/۸۳	۶۲/۰۵
	Recall	۶۸/۲۷	۸۲/۵۷	۸۲/۵۷	۸۳/۴۴	۸۷/۱۵
	F1-score	۳۹/۹۷	۶۳/۲۲	۶۳/۲۲	۵۹/۹۲	۷۰/۶۲
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{hog3d}$	Precision	۳۸/۱۳	۶۷/۰۵	۶۷/۰۶	۶۲/۹۲	**۷۶/۸۵**
	Recall	۷۴/۴۹	۸۹/۱۹	۸۹/۱۹	۸۹/۱۵	**۹۳/۰۵**
	F1-score	۴۸/۹۲	۷۴/۸۲	۷۴/۸۳	۷۲/۰۳	**۸۲/۹۳**
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{polygon}$	Precision	۳۹/۸۲	۶۷/۸۸	۶۷/۸۸	۶۳/۱۹	۷۲/۵۴
	Recall	۷۲/۱۰	۸۸/۴۱	۸۸/۴۱	۹۰/۳۳	۸۹/۹۷
	F1-score	۴۹/۱۴	۷۴/۸۵	۷۴/۸۵	۷۲/۵۳	۷۸/۸۴
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{polygon} + FV_{hog3d}$	Precision	۳۷/۴۳	۶۷/۱۴	۶۷/۱۴	۶۳/۲۰	۷۴/۲۸
	Recall	۷۲/۷۵	۸۹/۱۷	۸۹/۱۷	۸۹/۲۵	۹۱/۳۹
	F1-score	۴۷/۹۳	۷۴/۸۵	۷۴/۸۵	۷۲/۲۶	۸۰/۵۷
DCT [12]	Precision	۴۱/۲۳	۶۶/۸۰	۶۳/۷۶	۴۵/۵۲	۶۰/۳۳
	Recall	۶۶/۸۵	۸۴/۱۸	۸۲/۴۶	۶۹/۵۹	۸۲/۲۱
	F1-score	۳۱/۶۳	۷۲/۶۲	۷۰/۰۲	۵۳/۱۲	۶۷/۷۱
HOG [1]	Precision	۲۴/۷۸	۶۱/۹۳	۶۲/۵۴	۴۴/۴۳	۵۵/۷۶
	Recall	۶۰/۰۷	۸۵/۲۷	۸۴/۸۰	۷۲/۹۵	۷۸/۶۰
	F1-score	۳۳/۵۹	۶۹/۵۶	۶۹/۶۵	۵۲/۸۷	۶۳/۱۸
LBP [11]	Precision	۲۸/۵۴	۶۳/۳۹	۶۶/۷۶	۴۴/۸۸	۶۶/۴۹
	Recall	۶۷/۱۰	۸۵/۰۲	۸۷/۳۴	۷۱/۸۰	۸۸/۲۶
	F1-score	۳۸/۸۰	۷۰/۱۷	۷۳/۲۳	۵۲/۸۴	۷۳/۸۰
MBH [1]	Precision	۱۰/۸۸	۳۸/۱۸	۴۱/۰۱	۲۱/۰۹	۳۰/۰۵
	Recall	۳۹/۶۸	۷۴/۹۰	۷۶/۲۵	۵۷/۱۷	۶۵/۶۴
	F1-score	۱۶/۸۶	۴۸/۸۶	۵۱/۴۶	۲۹/۹۷	۳۹/۷۶
MBH + HOG [1]	Precision	۲۲/۹۱	۶۰/۷۰	۶۱/۰۲	۳۲/۱۹	۵۶/۵۰
	Recall	۵۶/۶۹	۸۶/۳۷	۸۶/۴۶	۶۷/۵۱	۸۰/۲۱
	F1-score	۳۱/۵۷	۶۹/۱۳	۶۹/۳۴	۴۲/۲۴	۶۴/۱۱

جدول ۴: نتایج تشخیص گفتار بر روی جملات و با آزمون وابسته به گوینده

ویژگی‌ها	DT	Linear SVM	Cubic SVM	KNN	RF
$FV_{gradient}^{spatial}$	۵۶/۵۰	۸۸/۸۰	**۹۴/۹۰**	۹۰/۳۰	۹۳/۵۰
$FV_{gradient}^{temporal}$	۳۳/۵۰	۷۹/۵۰	۸۶/۱۰	۸۸/۴۰	۸۱/۹۰
FV_{hog3d}	۴۲/۸۰	۸۵/۲۰	۸۶/۱۰	۷۷/۳۰	۶۹/۹۰
$FV_{polygon}$	۳۳/۴۰	۴۸/۰۰	۵۳/۱۰	۴۴/۱۰	۴۶/۵۰
$FV_{gradient}^{spatial} + FV_{polygon}$	۵۵/۲۰	۸۹/۳۰	۹۴/۵۰	۸۷/۷۰	۹۳/۱۰
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal}$	۵۳/۷۰	۹۰/۱۰	۹۳/۳۰	۸۸/۷۰	۹۳/۳۰
$FV_{gradient}^{spatial} + FV_{hog3d}$	۵۱/۴۰	۹۰/۴۰	۹۲/۱۰	۸۰/۴۰	۹۲/۴۰
$FV_{gradient}^{temporal} + FV_{hog3d}$	۴۰/۹۰	۸۸/۱۰	۸۹/۷۰	۸۱/۸۰	۸۸/۰۰
$FV_{hog3d} + FV_{polygon}$	۴۲/۹۰	۸۴/۸۰	۸۶/۶۰	۷۷/۷۰	۸۳/۵۰
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{hog3d}$	۵۳/۸۰	۹۱/۳۰	۹۲/۵۰	۸۲/۷۰	۹۲/۹۰
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{polygon}$	۵۴/۷۰	۸۹/۷۰	۹۳/۳۰	۸۹/۳۰	۹۳/۱۰
$FV_{gradient}^{spatial} + FV_{gradient}^{temporal} + FV_{polygon} + FV_{hog3d}$	۵۲/۴۰	۹۰/۹۰	۹۲/۵۰	۸۳/۹۰	۹۱/۹۰
DCT [12]	۵۱/۸۰	۸۸/۵۰	۹۳/۵۰	۹۰/۳۰	-
HOG [1]	۳۴/۹۰	۸۹/۲۰	۹۳/۶۰	۹۱/۲۰	-
LBP [11]	۳۵/۶۰	۸۳/۹۰	۸۸/۱۰	۹۰	-
MBH [1]	۲۷/۷۰	۸۳/۷۰	۸۷/۶۰	۹۰/۵۰	-
MBH + HOG [1]	۳۳/۶۰	۹۰/۸۰	۹۴/۱۰	۹۳	-

۳-۳- تأثیر اندازه همسایگی در روش گرادیان در نقاط کلیدی

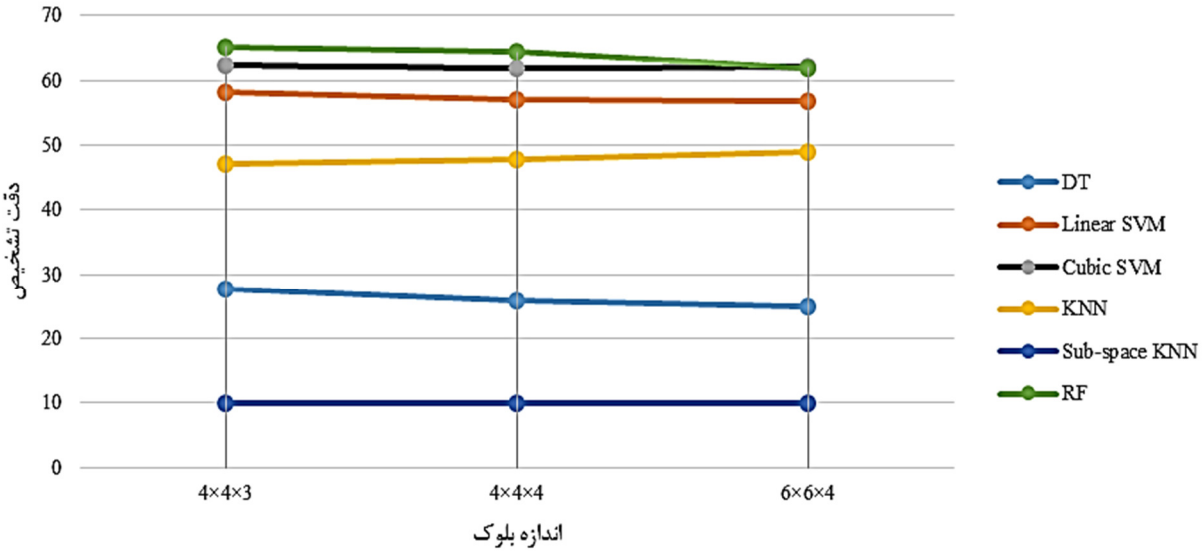
به‌منظور دست یافتن به‌اندازه همسایه مناسب که دقت تشخیص را افزایش دهد، تأثیر اندازه‌ی هم‌سایگی را برر سی خواهیم کرد. برای این کار ۵ سلول با اندازه‌های ۹×۹، ۷×۷، ۵×۵، ۳×۳ و ۱۶×۱۶ انتخاب شدند. تصویر ۹ نتایج آزمایش را با تغییر اندازه همسایگی نشان می دهد. از این شکل پیداست که بهترین عملکرد با همسایگی ۹×۹ پیکسل حاصل می‌شود؛ بنابراین برای آزمایش‌های بعدی همسایگی ۹×۹ استفاده شد.



تصویر ۹: بررسی تأثیر اندازه همسایگی در روش گرادیان در نقاط کلیدی لب

۴-۳- تأثیر اندازه بلوک انتخابی در HOG3D

به هنگام استفاده از ویژگی HOG3D اندازه بلوک انتخابی پارامتر مهمی است و در نتایج به‌دست‌آمده تأثیرگذار بوده، درنتیجه برای بررسی تأثیر اندازهٔ بلوک انتخابی ۴ بلوک مکعبی با اندازه‌های ۳×۴×۴، ۴×۴×۴ و ۴×۶×۶ انتخاب شد. تصویر ۱۰ نتایج آزمایش را با اندازه بلوک مکعبی نشان می‌دهد. از این شکل پیداست که بهترین عملکرد با اندازه بلوک مکعبی ۴×۶×۶ پیکسل حاصل می‌شود؛ بنابراین برای آزمایش‌های بعدی همسایگی ۴×۶×۶ استفاده شد.



تصویر ۱۰: بررسی تأثیر اندازه بلوک انتخابی در روش HOG3D

۴- نتیجه‌گیری

در این مقاله روشی برای تشخیص گفتار به کمک اطلاعات بصری با توصیف تغییرات مکانی-زمانی ناحیه لب ارائه شد. با توجه به این که در این روش لازم است فریم‌های گفتار بیان شده موجود باشد و با توجه به محدودیت‌های سخت‌افزاری موجود این روش برای تشخیص آنلاین قابل استفاده نیست. در روش پیشنهادی، پس از تشخیص صورت و تعیین ناحیه لب، نقاط کلیدی اطراف لب استخراج شد. گرادیان در نواحی مربوط به نقاط کلیدی اعمال‌شده و به عنوان اطلاعات مکانی مورد استفاده قرار گرفت. برای توصیف نواحی کلیدی لب در طول بیان یک عبارت، نمودار فراوانی ۳ بعدی گرادیان‌ها و تخمین مسیر تغییرات نواحی کلیدی در طول ویدیو استفاده شدند. در مرحله نهایی با استفاده از دسته‌بند مناسب، کلمه یا جمله بیان‌شده در ویدیو ورودی تشخیص داده شد. برای ارزیابی روش پیشنهادی از بانک داده MIRACL-VC1 استفاده شد. نتایج ارزیابی و مقایسه آن با روش‌های پیشین نشان‌دهنده برتری روش پیشنهادی بوده به میزان ۱۱ تا ۱۷ درصد بوده و همچنین نشان می‌دهد استفاده از ویژگی‌هایی که تغییرات لب‌ها را در طول زمان استخراج می‌کنند، به تشخیص بهتر گفتار بیان‌شده کمک می‌کند.

مراجع

[1] A. Rekik, A. Ben-Hamadou and W. Mahdi, “An adaptive approach for lip-reading using image and depth data,” Multimedia Tools and Applications, vol. 75, no. 14, pp. 8609-8636, 2016.

[2] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 198-213, 2002.

[3] K. Paleček, “Lipreading using spatiotemporal histogram of oriented gradients.” 24th European Signal Processing Conference, pp. 1882-1885, Aug. 2016.

[4] J. Shin, J. Lee and D. Kim, “Real-time lip reading system for isolated Korean word recognition,” Pattern Recognition, vol. 44, no. 3, pp. 559-571, 2011.

[5] G. Sterpu and N. Harte, “Towards lipreading sentences with active appearance models,” arXiv preprint arXiv:1805.11688, 2018.

[6] H. L. Bear, S. J. Cox and R. W. Harvey, “Speaker-independent machine lip-reading with speaker-dependent viseme classifiers,” arXiv preprint arXiv:1710.01122, 2017.

[7] P. Dalka, P. Bratoszewski and A. Czyzewski, “Visual lip contour detection for the purpose of speech recognition.” International Conference on Signals and Electronic Systems, pp. 1-4, Sept 2014.

[8] X. Ma, L. Yan and Q. Zhong, “Lip feature extraction based on improved jumping-snake model.” 35th Chinese Control Conference, pp. 6928-6933, 2016.

[9] F. Faridah and B. Achmad, “Lip image feature extraction utilizing snake's control points for lip reading applications,” International Journal of Electrical and Computer Engineering, vol. 5, no. 4, pp. 720, 2015.

[۱۰] نصیبه اسدی‌پرور ماسوله و اسدالله شاه‌بهرامی، «تخمین خودکار سن از روی تصویر چهره با تلفیق ویژگی‌های آماری و بافت»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۷، شماره ۳، صفحه ۸۲۹-۸۴۲، ۲۰۱۷.

[11] Y. Pei, T.-K. Kim and H. Zha, "Unsupervised random forest manifold alignment for lipreading," The IEEE International Conference on Computer Vision, pp. 129-136, 2013.

[12] A. Jain and G. Rathna, "Visual speech recognition for isolated digits using discrete cosine transform and local binary pattern feature," IEEE Global Conference on Signal and Information Processing, pp. 368-372, 2017.

[13] L. D. Terissi, M. Parodi, and J. C. Gómez, "Lip reading using wavelet-based features and random forests classification." 22nd International Conference on Pattern Recognition, pp. 791-796, 2014.

[14] S. S. Morade and S. Patnaik, "Lip reading by using 3-D discrete wavelet transform with dmey wavelet," International Journal of Image Processing (IJIP), vol. 8, no. 5, pp. 384, 2014.

[15] S .S. Morade and S. Patnaik, "Lip reading using DWT and LSDA," IEEE International Advance Computing Conference, pp. 1013-1018, 2014.

[۱۶] ساناز کشوری و عبدالله چاله‌چاله، «طبقه‌بندی سبک نقاشی هنرمندان با استفاده از هیستوگرام گرادیان جهت‌دار و الگوی باینری محلی»، مجله مهندسی برق دانشگاه تبریز، دوره۴۷، شماره ۳، صفحه ۱۱۹۵-۱۲۰۴، ۲۰۱۷.

[۱۷] منیره کوشش و غلامرضا اکبری‌زاده، «الگوریتم حذف Speckle با قابلیت حفظ لبه برای تصاویر سنجش‌ازدور رادار روزنه ترکیبی با استفاده از تبدیل چندمقیاسه‌ی Curvelet و آستانه‌گذاری وفقی»، مجله مهندسی برق دانشگاه تبریز، شماره ۴، دوره ۴۵، صفحه ۱۵۳-۱۶۱، ۲۰۱۵.

[18] G. Zhao, M. Barnard and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," IEEE Transactions on Multimedia, vol. 11, no. 7, pp. 1254-1265, 2009.

[19] W. C. Yau, D. K .Kumar and S. P. Arjunan, "Visual speech recognition using dynamic features and support vector machines," International Journal of Image and Graphics, vol. 8, no. 03, pp. 419-437, 2008.

[20] A. Rekik, A. Ben-Hamadou and W. Mahdi, "A new visual speech recognition approach for RGB-D cameras." International Conference Image Analysis and Recognition, pp. 21-28, 2014.

[21] A. Klaser, M. Marszałek and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients." 19th British Machine Vision Conference, pp. 275: 1-10, 2008.

[22] P. Viola and M. J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, no. 2, pp. 137-154, 2004.

[23] A. Asthana, S. Zafeiriou, S. Cheng and M. Pantic, "Incremental face alignment in the wild." The IEEE Conference on Computer Vision and Pattern Recognition, pp. 1859-1866, 2014.

[24] J. Fan, *Local Polynomial Modelling and its Applications: Monographs on Statistics and Applied Probability 66*: Routledge, 201.

زیرنویس‌ها

[1] Active Contour Model
[2] Active Shape Model
[3] Active Appearance Model
[4] Snake
[5] Local Binary Pattern
[6] Discrete Cosine Transform
[7] Histogram of Oriented Gradients
[8] Discrete Wavelet Transform
[9] Motion Boundary Histogram
[10] Local Binary Pattern from Three Orthogonal Planes
[11] Motion History Image
[12] Optical flow
[13] Integral Image
[14] Haar
[15] CHEHRA
[16] Support-vector machine
[17] k-nearest neighbors
[18] Decision tree
[19] Random forest
[20] Hyper plane
[21] kernel
[22] Precision
[23] Accuracy
[24] Recall
[25] F measure