
بسم الله الرحمن الرحيم

دانشکده برق و رباتیک

پایان نامه کارشناسی ارشد مهندسی برق گرایش الکترونیک

عنوان :

استخراج ویژگی زمانی- فرکانسی جهت شناسایی دیداری مصوت های فارسی

نگارش:

نسرین یادگار خسرویه

استاد راهنما:

آقای دکتر مروی

استاد مشاور:

آقای دکتر احمدی فرد

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

بهمن۱۳۹۲

شماره : ۱۱۳۵ /آ.ت.ب
تاریخ : ۹۲/۱۱/۲۷
ویرایش : ------

بسمه تعالی

مدیریت تحصیلات تکمیلی
فرم شماره (۶)

فرم صورتجلسه دفاع پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای :

نسرین یادگار خسرویه رشته : **بـــرق** گرایش : **الکترونیک (سیستم)**

تحت عنوان : **استخراج ویژگی زمانی- فرکانسی جهت شناسایی دیداری مصوتهای فارسی**

که در تاریخ **۹۲/۱۱/۲۷** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه : بسیار خوب امتیاز ۱۸/۵۵) ☑ دفاع مجدد ☐ مردود ☐

۱ـ عالی (۲۰ ـ ۱۹)
۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸) ✓
۳ـ خوب (۱۷/۹۹ ـ ۱۶)
۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)
۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضـــاء
۱ـ استاد راهنما		استادیار	
۲ـ استاد مشاور		استادیار	
۳ـ نماینده شورای تحصیلات تکمیلی		استادیار	
۴ـ استاد ممتحن		استادیار	
۵ ـ استاد ممتحن		استادیار	

رئیس دانشکده :

تقدیم به مادرم

که دست هایش می شویند غبار خستگی روزگار را و

سیراب می کنند روح تشنه ام را

تقدیم به روح پدرم

که دست هایش می تابانید نیرو را و محکم می کرد
استواری پایه های زندگیمان را

تقدیم به همسرم

آنکه آفتاب مهرش درآستانه قلبم همچنان پابرجاست

وهرگزغروب نخواهد کرد.

من لم یشکر المخلوق لم یشکر الخالق

لحظات می‌گذرد و عمر در گذر است. آنچه می‌ماند خاطرات خوش و کلام دلنشین استاد است که راه سعادت و نیکبختی را می آموزد چگونه می‌توان ایثار شمع و هستی بخشی او را به پروانه فقط در قالب الفاظ گنجانید؟ در وصف استاد سخن بسیار است اما آیا این همه الفاظ می‌تواند گویای جبران آن همه رنج و تلاش باشد. اینک به شکرانه خداوند متعال و با تلاش و راهنمایی خالصانه استاد ارجمندم آقای دکتر حسین مروی بر آن شدم تا گام های موثری در راه کسب علم و خدمت به جامعه بردارم. جا دارد از آن همه خلوص و راهنمایی صادقانه شما در انجام پایان نامه کمال تشکر و قدردانی را بنمایم. از جناب آقای دکتر احمدی فرد به خاطر راهنمایی ها و مساعدت هایشان در سمت استاد مشاور، کمال تشکر و قدرانی را دارم. و نیز از خانم وحیده السادات صادقی که پایگاه داده خود را در اختیار من قرار دادند کمال تشکر را دارم.

از عزیزترین افراد زندگیم خانواده‌ام و خانواده همسرم عاشقانه و از صمیم قلب سپاسگذارم. دعای خیرشان همیشه همراه من بوده و هر آنچه هستم، و هرآنچه دارم از برکت وجود آنهاست.

تعهد نامه

اینجانب **نسرین یادگار خسرویه** دانشجوی دوره کارشناسی ارشد رشته **الکترونیک (سیستم)** دانشکده **برق و رباتیک** دانشگاه صنعتی شاهرود نویسنده پایان نامه **استخراج ویژگی زمانی- فرکانسی جهت شناسایی دیداری مصوت‌های فارسی** تحت راهنمایی **آقای دکتر حسین مروی** متعهد می‌شوم.

تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.

- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « **Shahrood University of Technology** » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ : ۱۳۹۲/۱۱/۲۷

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

در این پایان‌نامه روشی برای شناسایی مصوت‌های فارسی در کلمات تک سیلابی ارائه می‌شود. برای این منظور پس از جداسازی فریم‌های تصویر و انتخاب فریم‌هایی که مربوط به تلفظ مصوت موجود در کلمه تک سیلابی بودند و نیز استخراج ناحیه‌ای پیرامون لب‌ها، ویژگی‌های مختلفی همچون ضرایب کسینوسی و ضرایب موجک و ضرایب MFCC برای تشخیص مصوت‌ها در کلمات تک سیلابی استخراج گردید. پس از آن توسط روش کاهش ویژگی LSDA، ویژگی‌ها را کاهش داده و سایز ویژگی‌ها را به ۲۵ تغییر دادیم. در نهایت موثرترین ویژگی‌ها برای شناسایی مشخص گردید. در این تحقیق از پایگاه داده‌ای شامل کلمات تک سیلابی، که توسط گویندگان مختلفی ادا شده بود و شامل ۵۸۰ ویدیو بود استفاده گردید. از ۳۸۱ ویدیو برای آموزش و از ۱۹۹ ویدیو برای آزمایش استفاده نمودیم. ویژگی‌های استخراجی به عنوان ورودی به شبکه عصبی دو لایه با ۲۰ نرون در لایه میانی و یک نرون در خروجی اعمال شدند. از تابع فعالسازی تانژانت سیگموید در لایه میانی و تابع خطی در خروجی استفاده کردیم و برای آموزش شبکه از روش گرادیان نزولی با نرخ آموزش متغیر استفاده نمودیم. بهترین نرخ شناسایی ۹۵.۷۵ بود که از محاسبه ضرایب MFCC از ۱/۴ بردار ضرایب DCT بعد از اسکن زیگزاگ ماتریس ضرایب کسینوسی به دست آمد.

کلمات کلیدی:

لب خوانی، شناسایی مصوت، ویژگی های زمانی- فرکانسی، کاهش ابعاد ویژگی، شبکه های عصبی

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

فصل اول : مقدمه

۱-۱ مقدمه

از دیر باز بشر، با این واقعیت آشنا بوده است که برای درک بهتر گفتار می‌تواند به حرکات لب و دهان گوینده در حین گفتار و هنگام ادای کلمات توجه کند. احتمالاً همه ما به طور ناخودآگاه تا حدی از این جنبه غیر صوتی گفتار استفاده کرده و هنگامی که محیط شنوایی، دچار همهمه و سر و صدا و آغشته به نویز صوتی می‌شود، به حرکات لب گوینده توجه بیشتری می‌کنیم. این امر در مورد مخاطبینی که دارای نقص در سیستم شنوایی خود هستند از اهمیت بالاتری برخوردار می‌باشد. ضمناً حرکات لب یا سیگنال تصویری گفتار می‌تواند به طور قابل ملاحظه‌ای دقت سیستم‌های تشخیص گفتار صوتی را خصوصاً در محیط‌های نویزی بهبود بخشد. همزمان کردن حرکات لب و صدای گفتار، برطرف کردن خطای تأخیر بین صوت و تصویر و دوبله اتوماتیک تصویری از دیگر کاربردهای این مقوله می‌باشد.

افرادی زیادی هستند که دچار آسیب در سیستم صوتی بوده و به دلیل عدم برخورداری از صدای مناسب، قادر به برقراری ارتباط با دیگران نیستند این افراد معمولاً توانایی انجام صحیح حرکات لب به شکلی که برای تکلم لازم است را داشته و در حالت ایده‌آل می‌توان با انجام لب‌خوانی به مقصود آن‌ها پی برد. گفتار بشری به دفعات به صورت صوتی و تصویری در طبیعت تکرار شده است. گفتار صوتی به شکل موج تولید شده توسط گوینده و گفتار دیداری به حرکات لب و زبان و ماهیچه‌هایی که در صورت است اشاره دارد. در گفتار صوتی واحد اصلی واج[1] نامیده می‌شود. در حوزه تصویری واحد اصلی از حرکات دهان ویزم[2] نامیده می‌شود که کوچک‌ترین جزء دیداری صحبت است. بسیاری از صداهای صوتی هستند که از نظر دیداری مبهم هستند این صداها به کلاس مشابه‌ای گروه‌بندی شده که یک ویزم را نشان می‌دهد. یک نگاشت چند به یک بین واج‌ها و ویزم‌ها هست یعنی می‌توان

[1] phonem
[2] viseme

مجموعه‌ای از واج‌ها را در نظر گرفت که تأثیر مشابه‌ای بر روی شکل دهان دارند. در جدول‌های زیر گروه‌بندی ویزم‌ها در زبان انگلیسی و فارسی آورده شده است [1] , [2].

جدول ۱- ۱ گروه‌بندی ویزم‌ها در انگلیسی

۱	p,b,m	۸	n,l
۲	f,v	۹	R
۳	th,dh	۱۰	A
۴	t,d	۱۱	E
۵	k,g	۱۲	I
۶	sh,zh	۱۳	O
۷	s,z	۱۴	U

جدول ۱- ۲ گروه‌بندی ویزم‌ها در زبان فارسی

۹. آ	۵. ر	۱. ف، و
۱۰. إ	۶.ج، چ، گ، ک، ن، ت، د، ی، ط	۲. ث، س، ص، ز، ذ، ظ، ض
۱۱. اً	۷. ای	۳. ژ، ش
۱۲. او	۸. أ	۴. ب، پ، م

به طور کلی سه روش برای شناسایی صحبت وجود دارد شامل شناسایی صوتی صحبت[1]، شناسایی تصویری صحبت[2]، شناسایی صوتی و تصویری صحبت[3]، که در این تحقیق به شناسایی تصویری صحبت پرداخته می‌شود.

۱-۲ ساختار پایان نامه

در فصل‌های مختلف این پایان نامه روش‌های شناسایی دیداری صحبت بررسی شده است. در فصل اول مقدمه‌ای در مورد شناسایی گفتار بیان شد. در فصل دوم به بررسی تحقیقات انجام شده در زمینه شناسایی دیداری صحبت و روش‌های مختلف برای انجام این کار پرداخته شده است. در فصل سوم روش‌های مختلف جداسازی دهان از بقیه قسمت‌های صورت معرفی شده است تا با استفاده از این روش‌ها بتوانیم علاوه برکوچک نمودن اندازه تصاویر، از پیچیدگی و نیز ابعاد زیاد ویژگی‌ها جلوگیری نماییم. در فصل چهارم نحوه محاسبه و استخراج ویژگی‌های فرکانسی - زمانی از ناحیه مورد نظر از دهان از فریم‌های مختلف ویدیو و نیز عملکرد آن‌ها با تغییر تعداد فریم‌های انتخابی و سایز تصاویر با یکی از روش‌های کاهش ویژگی نیز بررسی شده است. که این ویژگی‌های استخراجی برای تشخیص به شبکه عصبی اعمال شده‌اند و همچنین پایگاه داده‌ای که ما در این تحقیق از آن استفاده نمودیم معرفی شده است.

[1] Audio Speech Recognition
[2] Visual Speech Recognition
[3] Audio-Visual Speech Recognition

فصل دوم : مروری بر تحقیقات انجام شده

۲- ۱ مقدمه

شناسایی تصویری صحبت یا به عبارتی دیگر، لب خوانی شامل دو قسمت می‌باشد ابتدا استخراج ویژگی از تصاویر لب و سپس طبقه‌بندی (کلاسه‌بندی) ویژگی‌ها می‌باشد. برای استخراج ویژگی‌های تصویری دو روش مبتنی بر تصویر و مبتنی بر مدل را می‌توان استفاده نمود. در روش مبتنی بر تصویر ویژگی‌ها به طور مستقیم با اعمال تبدیل‌های ریاضی مانند تبدیل فوریه[۱]، تبدیل موجک[۲]، تبدیل کسینوسی گسسته[۳]، آنالیز مؤلفه‌های خاص[۴]، آنالیز مجزا ساز خطی[۵] بر روی تصاویر استخراج می‌شوند. مشکل این روش‌ها، ابعاد بزرگ و تکراری بودن داده‌ها و حساس بودن به چرخش و جابه‌جایی لب است. در روش مبتنی بر مدل، مدلی از لب ساخته شده و به وسیله مجموعه کوچکی از پارامترها توصیف می‌شود همچون مدل‌های شکل فعال[۶]، مدل‌های مرز فعال[۷]، الگوهای انعطاف پذیر[۸]، که مزیت مزیت این روش، بیان ویژگی‌ها در ابعاد کوچک و تأثیر ناپذیری مدل از روشنایی تصویر، چرخش، اندازه و جابه‌جایی لب است.

۲-۲ مدل‌های مرز فعال

یکی از روش های مبتنی بر مدل که روش بالا به پایین نیز نامیده می شوند مدل کانتور فعال می باشد. پتاجان[۹] احتمالاً اولین محقق برای توسعه سیستم لب خوانی بوده است [3]. مدل مرز فعال توسط منحنی باز یا بسته با تعدادی نقاط کنترل نزدیک تصویر شی‌ای که می‌خواهیم شکل آن را استخراج کنیم مدل می‌شود. برای فرم‌پذیری آن چند فاکتور انرژی در نظر گرفته می‌شود و با کمینه کردن این

[1] Fourier Transform
[2] Wavelet Transform
[3] Discrete Cosine Transform
[4] Principal Component Analysis
[5] Linear Discriminant Analysis
[6] Active Shape Models
[7] Active Contour Models
[8] Deformable Templates
[9] Petajan

انرژی‌ها منحنی فرم لازم را به خود می‌گیرد. این مدل توسط گس و همکارانش معرفی شد [4] که به دلیل شباهت حرکت کانتور[1] به خزش مار[2]، آن‌ها این مدل را مار نامیدند. مار می‌تواند توسط تعدادی نقطه، انرژی کشسان داخلی[3] و یا انرژی بر اساس لبه خارجی بیان شود.

۲-۲-۱ تابع انرژی

یک مار می‌تواند توسط n نقطه به صورت $V_i= (x_i , y_i)$, i=0, 1, 2, …., n-1 نمایش داده شود.

تابع انرژی مار به صورت زیر بیان می‌شود.

$$E^*_{snake}= \int_0^1 E_{snske}\,(V(s))\,ds= \int_0^1 (E_{internal}\,(V(s)) + E_{image}\,(V(s)) + E_{con}(V(s)))ds$$

رابطه (۲- ۱)

$$E_{external} = E_{image} + E_{con}$$ رابطه (۲- ۲)

$$E_{internal} = E_{cont} + E_{curv}$$ رابطه (۲- ۳)

که انرژی خارجی از مجموع انرژی تصویر و انرژی محدودیت خارجی[4] که توسط کاربر اعمال می‌شود تشکیل شده است. انرژی داخلی مجموع انرژی کانتور مار و انرژی خمش مار[5] می‌باشد.

$$E_{internal} = (\alpha(s)|Vs(s)|^2 + \beta(s)|V_{ss}(s)|^2) /2$$

$$= (\alpha(s) \,||\, d\overline{v}(s) \,||^2 +\beta(s) \,||d^2\,\overline{v}(s)||^2)/2$$ رابطه (۲- ۴)

[1] Contour
[2] Snake
[3] Elastic
[4] External Constrain
[5] Curvature

مقادیر بزرگ $\alpha(s)$ و $\beta(s)$ انرژی داخلی مار را هنگامی‌که خیلی زیاد گسترش می‌یابد افزایش خواهد داد و مقادیر کوچک آن‌ها محدودیت‌های کمتری روی اندازه و شکل مار قرار می‌دهند.

نیروی تصویر شامل سه مؤلفه انرژی، انرژی خطوط، انرژی لبه‌ها، انرژی ختم شدگی‌ها می‌باشد.

$$E_{image} = w_{line} E_{line} + w_{edge} E_{edge} + w_{term} E_{term}$$ رابطه (۲- ۵)

که تنظیم وزن‌ها، ویژگی‌های برجسته تصویر را که توسط مار فرض شده مشخص می‌کند.

$$E_{line} = I(x, y)$$ رابطه (۲- ۶)

$$E_{edge} = - |\nabla I(x,y)|^2$$ رابطه (۲- ۷)

انرژی لبه را به صورت زیر می‌توان نوشت که G_σ یک گوسی با انحراف استاندارد σ می‌باشد.

$$E_{edge} = - | G_\sigma * \nabla^2 I|^2$$ رابطه (۲- ۸)

انحناء سطح خطوط در یک تصویر کمی یکنواخت شده برای مشخص کردن گوشه‌ها و ختم شدگی‌ها در تصویر استفاده می‌شود. فرض کنید $C(x,y)$ یک نسخه یکنواخت شده از تصویر باشد به طوری‌که

$$C(x, y) = G_\sigma * I(x, y)$$ رابطه (۲- ۹)

$$\theta = \arctan\left(\frac{C_y}{C_x}\right)$$

$$E_{term} = C_{yy} C_x^2 - 2C_{xy} C_x C_y + C_{xx} C_y^2 \,/\, (C_x^2 + C_y^2)^{3/2}$$ رابطه (۲- ۱۰)

انرژی ختم شدگی از رابطه (۲- ۱۰) به دست می‌آید.

۲-۲-۲ حداقل سازی انرژی

برای انطباق منحنی به کانتور باید انرژی حداقل (می‌نیمم) شود به همین دلیل با استفاده از یکی از روش‌های می‌نیمم سازی این کار باید انجام پذیرد. در این‌جا از روش شیب (گرادیان) نزولی[1] که از ساده‌ترین بهینه‌سازهاست استفاده شده که روابط آن در زیر بیان شده است.

$x_{t+1}= x_t + \gamma\, df(x_t)/dx$ رابطه (۲- ۱۱)

$y_{t+1} = y_t + \gamma\, df(y_t)/dy$ رابطه (۲- ۱۲)

γ، مقدار گام را در هر تکرار کنترل می‌کند.

$\overline{x}_{t+1}= \overline{x}_t + \gamma\, \nabla f(\overline{x}_t)$ رابطه (۲- ۱۳)

انرژی مار را به صورت مجموع انرژی نقاط گسسته روی مار می‌توان تقریب زد.

$E^*_{snake} \approx \sum_1^n E_{snake}(\overline{v}_i)$ رابطه (۲- ۱۴)

$\nabla E^*_{snake} \approx \sum_1^n \nabla E_{snake}(\overline{v}_i)$

$\overline{vi} \leftarrow \overline{vi} - \nabla E_{snake}(\overline{vi})$

معادله‌های نهایی در زیر آورده شده است.

$$\overline{vi} = \overline{vi} - \gamma\left\{ w_{internal}\left[\alpha\, \partial^2\overline{v} / \partial s^2 + \beta\, \partial^4\overline{v} / \partial s^4\right] + \nabla E_{ext}(\overline{vi})\right\}$$ رابطه (۲- ۱۵)

$$\overline{xi} = \overline{xi} - \gamma\left\{ w_{internal}\left[\alpha\, \partial^2 x / \partial s^2 + \beta\, \partial^4 x / \partial s^4\right] + \partial E_{ext}(\overline{vi}) / \partial x\right\}$$

[1] Gradient-descent

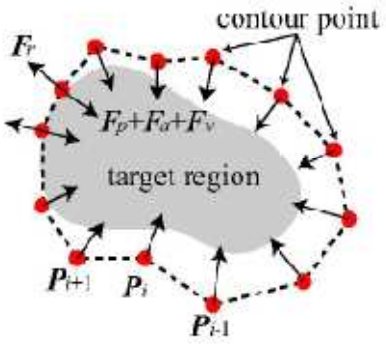
$$\overline{yi} = \overline{yi} - \gamma \left\{ w_{internal} \left[\alpha \, \partial^2 y / \partial s^2 + \beta \, \partial^4 y / \partial s^4 \right] + \partial E_{ext}(\overline{vi}) / \partial y \right\}$$

که در نهایت مقادیر نقاط روی مرز به دست خواهد آمد.

در [5] مدل کانتور فعال به کار گرفته شده که از مجموعه‌ای آموزشی شامل ۴۵۰۰ تصویر مربوط به حروف آلمانی که توسط ۶ شخص بیان شده استفاده شده است. ابتدا تصاویر با الگوریتم[1] متداول مار برچسب خورده‌اند و مارهای هم‌تراز نشده به صورت دستی از پایگاه داده خارج شده‌اند و هر کانتور لب به یک بردار ۸۰ بعدی که در واقع به صورت ۴۰ نقطه دو بعدی می‌باشد کد شده است. برای دنبال کردن[2] و یافتن لب در تصاویر جدید انرژی را که منفی مجموع تمام گرادیان‌های سطح خاکستری تخمین زده شده در طول کانتور می‌باشد محاسبه کرده‌اند. انرژی محلی می‌نیمم، تطبیقی از مدل کانتور با مرز واقعی لب را نشان می‌دهد. این عمل با استفاده از گرادیان نزولی صورت گرفته چون مرز خارجی لب‌ها ویژگی خیلی قوی برای دنبال کردن می‌باشد. بعد از یافتن کانتور لب "Eigenlips " محاسبه شده‌اند. n مولفه خاص اول کانتورها و یا n مولفه خاص اول از ماتریس تصویر سطح خاکستری پیرامون لب را انتخاب نموده که به آن‌ها "Eigenlips " می‌گویند. میانگین و بردارهای ویژه از لب‌ها به دست آورده شده است. ده مؤلفه خاص اول برای جداسازی تمام شکل‌های سطح خاکستری کافی است. ماتریس سطح خاکستری کدگذاری نسبت به جابه‌جایی، چرخش، مقیاس تغییر ناپذیر است اما نسبت به روشنایی تغییرپذیر است. با استفاده از طبقه‌بند (کلاسه‌بند) MLP[3] این تغییر پذیری می‌تواند برطرف شود چون فقط یکی از ده ویژگی شدیداً به روشنایی وابسته است. از ویژگی‌های صوتی و تصویری استفاده شده و به کلاسه‌بند MLP اعمال و احتمال پسین[4] به دست آورده شده است. مطابق با قانون بیز[5] احتمال‌ها به دست آمده و به عنوان احتمال گذر برای مدل‌های

[1] Algorithm
[2] Tracking
[3] Multi Layer Perceptron
[4] Posterior Probability
[5] Bayes Law

مخفی مارکوف[1] استفاده شده است. در [6] از مدل کانتور فعال نمونه‌گیری شده[2] برای آشکارسازی کانتور لب در دنباله‌های تصویر ورودی استفاده شده است. این مدل برای تصاویر دودویی به کاربرده شده است. این مدل حلقه‌ی بسته‌ای از چند ضلعی به وجود آمده توسط نقاط کانتور که با چهار نیرو کار می‌کند است که در شکل زیر نشان داده شده است.



شکل ۲- ۱ مدل کانتور فعال نمونه‌گیری شده

F_p نیروی فشار[3] که در جهت نیم ساز دو نقطه کنترل مجاور عمل می‌کند و مقداری ثابت است.

F_a نیروی کشش[4] که نسبت به فاصله دو نقطه کنترل مجاور عمل می‌کند.

F_v نیروی لرزش[5] که مقداری ثابت است و در جهت عمود بر برآیند دو نیروی قبل عمل می‌کند و جهت آن در هر حلقه معکوس می‌شود.

[1] Hidden Markov models
[2] Sampled Active Contour Model
[3] Pressure
[4] Attraction
[5] Vibration

F_r نیروی دفع[1] می‌باشد که هنگامی‌که نقطه کنترل به مرز شی می‌رسد این نیرو در جهت خلاف نیروهای دیگر عمل می‌کند.

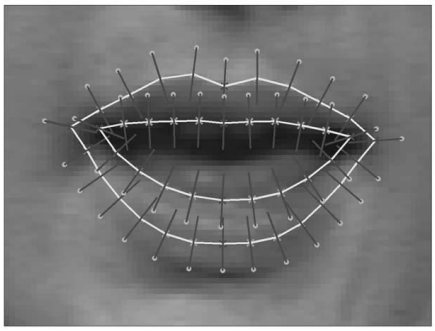
در [7] یک مدل فرم پذیر بر اساس کانتور فعال با چهار نوع انرژی برای نقاط کنترل در نظر گرفته شده است. با ترکیب مناسب این انرژی‌ها و کمینه کردن آن در دو مرحله برای استخراج لبه‌های قوی و ضعیف، شکل بیرونی دهان و لب‌ها و پارامترها استخراج می‌گردند. در این مقاله بعد از تخمین اولیه-ی محل دهان و اصلاح آن، در دو مرحله لبه بالا و پایین دهان استخراج می‌شود. به دلیل استخراج هر یک در مراحل جداگانه و نیز عدم نیاز به نزدیک بودن کانتور اولیه به لبه‌های استخراج شده نسبت به تغییرات شدت لبه بالا و پایین مقاوم است.

۲- ۳ مدل‌های شکل فعال

مدل شکل فعال مبتنی بر یک الگوریتم تطبیقی تکراری است که تحت تاثیر محدودیت‌های شکل قرار می‌گیرد. این محدودیت‌ها با توجه به مدل آماری شکل که به آن مدل توزیع نقطه‌ای[2] می‌گویند تعیین می‌گردند. که از آمار به دست آمده از اطلاعات داده‌های آموزشی که به صورت دستی نشانه‌گذاری شده‌اند به دست می‌آید. مدل توزیع نقطه‌ای کاهش فضای شکل‌های معتبر لب را در مفهوم داده آموزشی توصیف می‌کند و نقاط در این فضا نماینده‌های فشرده‌ای از شکل لب هستند که به صورت مستقیم می‌توانند استفاده شوند.

هر مدل شکل توسط مختصات نقاط مشخص شده نمایش داده می‌شود. در شکل زیر یک مدل لب با ۴۴ نقطه نشان داده شده است. (۲۴ نقطه روی کانتور خارجی و ۲۰ نقطه روی کانتور داخلی).

[1] Repulsion
[2] point Discriminate Model



شکل ۲- ۲ علامت‌گذاری انجام شده بر روی لب

ابتدا گوشه‌ها به صورت دستی تعیین و سپس بقیه نقاط با فاصله‌های یکسان بین آن‌ها قرار می‌گیرند. اگر i امین شکل مدل با $x_i = (x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{i44}, y_{i44})$ بیان شود دو شکل مشابه x_1 و x_2 توسط می‌نیمم سازی انرژی هم‌تراز[1] می‌شوند.

$$E = (x_1 - M(s,\theta)[x_2] - t)^T w(x_1 - M(s,\theta)[x_2] - t) \qquad \text{رابطه (۲- ۱۶)}$$

جایی که تبدیل مقیاس با s، چرخش با θ و جابه‌جایی در x , y با t_x , t_y نشان داده شده است.

$$M(s,\theta)\begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} + (s\cos\theta)y_{jk} \end{pmatrix} \qquad \text{رابطه (۲- ۱۷)}$$

$$t = (t_{x1}, t_{y1}, \ldots\ldots, t_{xN}, t_{yN})$$

w ماتریس وزن قطری در هر نقطه است که مقادیر وزن‌های آن در هر نقطه با واریانس آن نقطه نسبت عکس دارد. برای هم‌ترازی از الگوریتم تکراری بیان شده در [8] استفاده شده است. بنابراین مجموعه‌ای از مدل‌های شکل هم‌تراز شده به دست می‌آید، متوسط شکل $\overline{x_s}$ محاسبه شده و محورهایی که بیشترین واریانس را از شکل متوسط توصیف می‌کنند می‌توانند توسط آنالیز مؤلفه‌های خاص مشخص شوند. هر شکل می‌تواند توسط رابطه زیر تقریب زده شود.

$$x_s = \overline{x_s} + P_s b_s \qquad \text{رابطه (۲- ۱۸)}$$

[1] Alignment

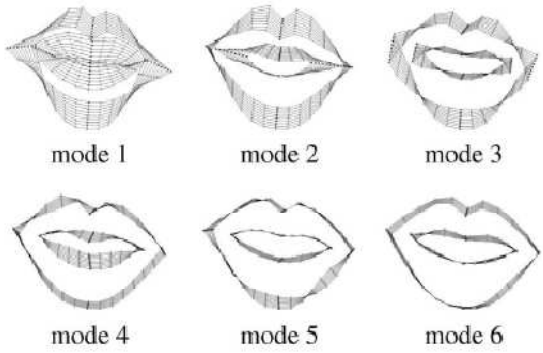
$P_s=(P_1 ,P_2 ,\ldots..,P_t)$ ماتریسی از اولین بردارهای ویژه است و b_s یک بردار از وزن‌های t است

$b_s = (b_1 ,b_2 ,\ldots..,b_t)$ چون بردارهای ویژه متعامدند پارامترهای شکل b_s می‌تواند به صورت زیر محاسبه شود.

$$b_s = P_s{}^T (x_s - \overline{x_s})$$ رابطه (۲- ۱۹)

این اجازه می‌دهد که شکل‌های معتبر لب به صورت فشرده نمایش داده شود. تعداد حالت‌های متغیر از تعداد نقاط علامت‌گذاری شده بسیار کمتر است.

۶ حالت از مدل توزیع نقطه‌ای از ۱۱۴۴ تصویر آموزشی از پایگاه داده Av Letters که به صورت دستی برچسب‌گذاری شده‌اند در شکل زیر نشان داده شده است.



شکل ۲- ۳ مدل توزیع نقطه‌ای، هر حالت با $2\sigma \pm$ اطراف متوسط رسم شده است

برای تطبیق تکراری مدل توزیع نقطه‌ای تابع هزینه مورد نیاز است. که این تابع هزینه باید می‌نیمم شود.

$$e = (g - g_{mean})^T (g - g_{mean}) - b_t{}^T b_t$$ رابطه (۲- ۲۰)

در تابع هزینه e، g پروفایل[1] سطح خاکستری ، g_{mean} میانگین بردار پروفایل سطح خاکستری است.

[1] Profile

$b_t = P^T (g - g_{mean})$ رابطه (۲- ۲۱)

پارامترها توسط b_t توصیف می‌شوند [9]. این روش همچنین در [10] برای استخراج پارامترهای شکل استفاده شده و به همراه شدت روشنایی به عنوان ویژگی‌های تصویری صحبت استفاده شده‌اند.

در [11] یک سیستم لب خوانی اتوماتیک با استفاده از اطلاعات دیداری برای شناسایی ارقام انگلیسی مجزا از صفر تا نه ارائه شده است که از یک مدل شکل فعال چهارده نقطه‌ای برای توصیف کانتور خارجی لب استفاده نموده است. که بعد از فرآیند بهینه‌سازی، مجموعه پارامترهای بهینه شامل

$\{ x_c, y_c, s, \theta, b_0 \}$ بدست می‌آید. که x_c, y_c نقطه مرکزی از مدل لب می‌باشد و s فاکتور مقیاس، θ زاویه چرخش و b_0 بردار وزن برای بردارهای ویژه است. که بردار وزن اطلاعات شکل را شامل می‌شود و برای تشخیص شکل‌های متفاوت دهان اهمیت اساسی دارد. چون تغییرات در s و θ به تنظیمات دوربین وابسته است این پارامترها نمی‌توانند به بهبود عملکرد شناسایی کمک کنند. بنابراین، این دو پارامتر نرمالیزه شده، که نسبت به مقادیر به دست آمده آن‌ها از تصویر اول در دنباله تصویر لب، مفیدتر واقع می‌شوند. از این رو، بردار ویژگی تصویری $\{ s_{normalized}, \theta_{normalized}, b_0 \}$ برای توصیف کانتور خارجی لب استفاده شده است.

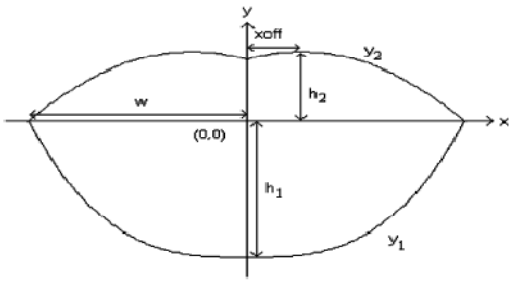
در [12] برای استخراج اطلاعات در مورد شکل و حرکت لب‌ها از مدل‌های شکل فعال استفاده شده است. مدلی که در این‌جا استفاده شده است شامل دو گروه اصلی اطلاعات سطح خاکستری و اطلاعات شکل می‌باشد. اطلاعات شکل برای پارامتری کردن صحبت و اطلاعات سطح خاکستری برای کمک به دنبال کردن لب‌ها استفاده شده است. مدل شکل فعال توسط مدلی از پروفایل سطح خاکستری اطراف کانتور لب، لب‌ها را دنبال می‌کند. از ۲۷ نقطه، با بردارهای پروفایل سطح خاکستری به طول ۹ که از هر نقطه می‌گذرد استفاده شده است. تصاویر از پایگاه داده TULIPS1 انتخاب شده‌اند. برای هر فریم، پارامترهای شکل و شدت روشنایی با مدل شکل فعال استخراج شده، مدل‌ها با ۲۰ پارامتر شکل و ۱۰

پارامتر شدت روشنایی آموزش داده شده‌اند. در [13] از مدل شکل فعال بر اساس منحنی استفاده شده است. که از ۵ منحنی سهمی شکل برای نمایش لب استفاده شده است. که برای نمایش این سهمی‌ها سه ضریب لازم است. در این روش نسبت به مدل توزیع نقطه‌ای پارامترهای کمتری مورد نیاز است.

۲-۴ مدل‌های انعطاف‌پذیر[1]

در این روش ابتدا یک مدل هندسی برای لب مشخص شده و سپس یک تابع انرژی که پارامترهای مدل را به مرزهای شکل مرتبط می‌کند تعریف می‌شود. این تابع میزان تطبیق بین مدل و مرزهای شکل را برای هر وضعیت اندازه‌گیری کرده و وضعیتی را که کمترین مقدار تابع انرژی را فراهم سازد به عنوان بهترین انطباق بر می‌گزینند. از این رو جستجویی در تصویر گرادیان و پارامترهای الگو انجام می‌شود تا شکل لب در هر تصویر تعیین شود. در فریم‌های بعدی از شکل و موقعیت مدل، در فریم-های قبلی استفاده شده و پارامترهای هندسی تشکیل دهنده الگو به عنوان مشخصه استخراج می‌شود. این روش ناحیه لب و غیر لب را بر اساس رنگ و شدت روشنایی توسط یک مدل لب هندسی ساده جدا می‌سازد[14].

۲-۴-۱ مدل لب



شکل ۲- ۴ مدل هندسی لب

[1] Deformable Models

یک مدل هندسی انعطاف‌پذیر برای لب در نظر می‌گیریم چون مدل هندسی اجازه می‌دهد که شکل لب توسط مجموعه کوچکی از پارامترها توصیف شود. معادلات مربوط به مدل شکل (۲- ۴) به شرح زیر است.

$$y_1 = h_1 \left(\left(\frac{x - sy_1}{w} \right)^2 \right)^{1+\delta^{\wedge}2} - h_1$$ رابطه (۲- ۲۲)

$$y_2 = \frac{-h2}{(w - x_{off})^2} \left(|x - sy_2| - x_{off} \right)^2 + h_2$$ رابطه (۲- ۲۳)

$x \in [-w, w]$ و $(0,0)$ مرکز می‌باشد. s انحراف شکل لب و δ انحراف منحنی y_2 از منحنی قائم را نشان می‌دهند. هنگامی‌که مرکز مدل در (x_c, y_c) قرار می‌گیرد و لب انحراف θ نسبت به مرکز مدل دارد.

x را با $(x - x_c)\cos\theta + (y - y_c)\sin\theta$ و y را با $-(x - x_c)\sin\theta + (y - y_c)\cos\theta$ جایگزین نموده‌اند.

در نتیجه مجموعه پارامترها که شکل لب را کنترل می‌کنند توسط مجموعه‌ای به صورت

$p = \{ x_c , y_c , w , h_1 , h_2 , x_{off} , \delta , s , \theta \}$ نشان داده شده‌اند.

۲- ۴-۲ فرمول‌بندی تابع هزینه[1]

هدف قطعه‌بندی تصویر به دو ناحیه لب و غیر لب می‌باشد. اگر به هر پیکسل در تصویر یک احتمال تعلق به پیکسل لب اختصاص داده شود سپس تابع هزینه که در ذیل آمده به معیار حداکثر (ماکزیمم) احتمال منجر می‌شود که می‌تواند برای مشخص نمودن بخش‌های پیش‌زمینه و پس‌زمینه استفاده شود.

$$C(p) = -\prod_{i=1}^{2} \prod_{(x,y)\epsilon R_i} prob_i(x, y)$$ رابطه (۲- ۲۴)

[1] Cost Function

که R_1 و R_2 به ترتیب ناحیه لب و غیر لب می‌باشند. $Prob_1(x , y)$ احتمال پیکسل در مکان (x , y) متعلق به پیکسل‌های لب است و $Prob_2(x , y)= 1- Prob_1(x , y)$ احتمال پیکسل در مکان (x , y) متعلق به پیکسل‌های غیر لب می‌باشد. λ پارامترهای مدل را تعیین می‌کند. با لگاریتم‌گیری و بسط به فضای پیوسته داریم:

$$E(p) = - \int_{x_1(p)}^{x_2(p)} \int_{y_1(p;x)}^{y_2(p;x)} g(x, y)dxdy$$ رابطه (۲- ۲۵)

$$g(x , y)= \log prob_1(x , y) - \log prob_2(x , y)$$ رابطه (۲- ۲۶)

که $x_1(p) = x_c - w \cos\theta$ و $x_2(p) = x_c + w \cos\theta$ نقاط گوشه چپ و راست لب هستند. $y_1(p;x)$ و $y_2(p;x)$ نقاط مرز عمودی از خط x هستند.

پارامترهای بهینه مدل تابع هزینه رابطه (۲- ۲۵) را می‌نیمم می‌کنند. در اینجا برای یافتن احتمال هر پیکسل متعلق به لب یا به ناحیه غیر لب از خوشه‌بندی فازی[1] استفاده شده است.

ناحیه بهینه هنگامی که رابطه (۲-۲۴) ماکزیمم شود به دست می‌آید. ماکزیمم بودن این رابطه با می-نیمم بودن رابطه (۲- ۲۵) معادل می‌باشد.

۲-۴-۳ بهینه‌سازی پارامترهای مدل

با استفاده از گرادیان نزولی تابع هزینه در رابطه (۲- ۲۵) می‌نیمم می‌شود. با مشتق گرفتن نسبت به پارامترهای مدل رابطه زیر حاصل شده است.

$$\frac{\partial E}{\partial p_m} = \int_{x_1(p)}^{x_2(P)} \left[g\big(x , y_1(p; x)\big) \frac{\partial y_1(p;x)}{\partial p_m} - g\left(x , y_2(p; x)\right) \frac{\partial y_2(p;x)}{\partial p_m} \right] dx$$

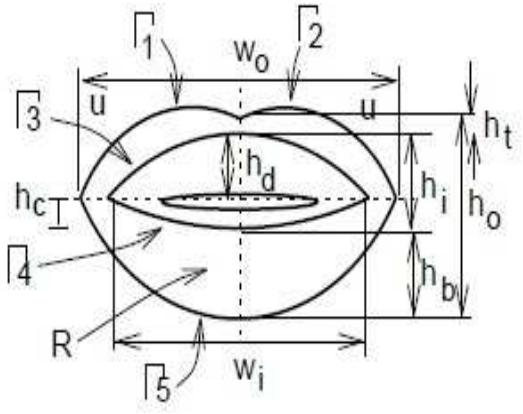
رابطه (۲- ۲۷)

[1] Fuzzy Clustering

که $p_1 = x_c$, $p_2 = y_c$, $p_3 = w$, ….., $p_8 = s$, $p_9 = \theta$ می‌باشند.

۲- ۵ الگوهای انعطاف پذیر

در [15] از الگوهای انعطاف پذیر برای مدل کردن لب استفاده شده است.



شکل ۲- ۵ الگوی لب

همان طور که در شکل بالا دیده می‌شود برای مدل کردن لب از سهمی و برای مدل کردن زبان از بیضی استفاده شده است. که معادلات مربوط به آن‌ها در ذیل آورده شده است.

رابطه (۲- ۲۸)

$$y_{\Gamma 1}(x) = (h_d + h_t)\left(1 - \frac{2x}{w_0}\right)^2 - ux - \frac{2u}{w_0}x^2 \; ; \; -w_0/2 < x < 0$$

$$y_{\Gamma 2}(x) = (h_d + h_t)\left(1 - \frac{2x}{w_0}\right)^2 + ux - \frac{2u}{w_0}x^2 \; ; \; 0 < x < w_0/2$$

$$y_{\Gamma 3}(x) = h_d\left(1 - \frac{2x}{w_i}\right)^2 ; \qquad -w_i/2 < x < w_i/2$$

$$y_{\Gamma 4}(x) = h_c\left(1 - \frac{2x}{w_i}\right)^2 ; \qquad -w_i/2 < x < w_i/2$$

$$y_{\Gamma 5}(x) = -(h_d + h_c)\left(1 - \frac{2x}{w_0}\right)^2 ; \quad -w_0/2 < x < w_0/2$$

ناحیه حفره‌ی دهانی بین لب پایینی و بالایی $2/3\ h_i w_i$ است. ناحیه لب‌ها به صورت

$R= 2/3\ h_0 w_0 + (\ uw_0 /12)\ w_0$ می‌باشد. پارامترها برای بیضی مدل شده برای زبان y_{ton} که مرکز عمودی از زبان، h_{ton} ارتفاع زبان، w_{ton} پهنای زبان هستند. مساحت قابل مشاهده از زبان $\pi /4\ h_{ton} w_{ton}$ است. برای سادگی تمام سهمی‌ها به یک کانتور $\Gamma \epsilon\ \partial R$ از الگوی انعطاف‌پذیر گروه‌بندی می‌شوند.

$$\vec{v}(\tilde{s}) = (\ x\ ,\ y_{\Gamma}\ (x)\ )^T = (x(\tilde{s})\ ,y(\tilde{s})\ )^T$$ رابطه (۲- ۲۹)

که $\tilde{s} \epsilon\ [0,1]$ و N گره که $n \epsilon\ \{۱,\ldots..,N\}$.

الگوی انعطاف‌پذیر سعی در می‌نیمم سازی انرژی دارد. انرژی‌ها می‌توانند وابسته به دره‌ها[1] یا قله‌های نواحی در تصویر تعریف شوند.

$$E_v = \frac{c}{|R|} \int \Phi_v\ (\vec{v})\ dA$$ رابطه (۲- ۳۰)

یا روی لبه‌های تصویر به شکل رابطه (۲- ۳۱) تعریف شود.

$$E_e = \sum_{k=1}^{5} \frac{c_x}{|\Gamma_k|} \int \Phi_e(\vec{v})\ d\tilde{s}$$ رابطه (۲- ۳۱)

$\Phi_e\ (\vec{v})\ ,\ \Phi_v\ (\vec{v})$ پتانسیل‌های لبه و دره‌ها از تصویر هستند.

انرژی محدودیت داخلی $E_{con}= k/2\ (w_0 - \lambda h_0)^2\ ;\ \lambda \approx$ ۲ می‌باشد. پارامترهایی همچون، $(\theta\ ,\ w_0\ ,\ h_0)$ توسط تابع انرژی E_v و بقیه پارامترها توسط می‌نیمم سازی انرژی لبه E_e تنظیم می‌شوند.

یکی دیگر از روش‌ها برای شناسایی دیداری صحبت روش مبتنی بر تصویر است که روش پایین به بالا نیز نامیده می‌شود. در واقع این روش مبتنی بر شدت روشنایی پیکسل‌های تصویر است و هر گونه پردازشی روی این مقادیر شدت روشنایی صورت می‌گیرد و ویژگی‌ها از آن‌ها استخراج می‌شود.

[1] Valley

۲- ۶ موجک هار[1]

یکی از روش‌های مبتنی بر تصویر DWT می‌باشد که کاربردهای زیادی در حوزه استخراج ویژگی به خصوص از تصویر دارد و یکی از ویژگی‌های مؤثر در شناسایی تصاویر می‌باشد. که مستقیما از خود این ضرایب استفاده شده یا از سایر ویژگی‌ها و مشخصه‌ها که از این ضرایب استخراج می‌شود استفاده شده است. از دیگر کاربردهای ویولت حذف نویز از سیگنال‌ها و تصاویر با حذف محدوده خاصی از ضرایب موجک و فشرده‌سازی تصاویر می‌باشد.

یکی از ساده‌ترین انواع این تبدیل ، ویولت هار می‌باشد که در [16] این روش به کار گرفته شده و لب با استفاده از این تبدیل قطعه‌بندی شده و بردارهای ویژگی از نتیجه آن استخراج می‌شود. قطعه‌بندی لب در چند مرحله صورت گرفته است.

۲- ۶- ۱ پیش پردازش

شرایط نوری برای نرمالیزه کردن سطح روشنایی در تصویر اصلی با تابع لگاریتمی به صورت زیر تغییر داده شده است.

$$g(x,y) = k + \frac{\log(f(x,y)+1)}{d*\log(t)}$$ رابطه (۲- ۳۲)

که $f(x , y)$ تصویر اصلی و $g(x,y)$ تصویر پیش پردازش شده است. مجموعه پارامترهای (k , d , t) برای کنترل موقعیت و شکل منحنی تنظیم شده‌اند. در این مطالعه به طور نسبی (۲، ۰.۵، ۱۲) در نظر گرفته شده‌اند.

[1] Haar Wavelet

۲ – ۶ – ۲ تبدیل رنگی

رنگ لب و ناحیه پوست معمولاً هم‌پوشانی دارند بنابراین فضای رنگی خاصی باید برای نشان دادن تغییرهای کوچک انتخاب شود. از آن‌جا که فاصله بین هر دو نقطه در فضای رنگی متناسب با تفاوت رنگ آن‌ها است. یک فضای رنگی یکنواخت نیاز است. تصویر رنگی به فضای رنگی $CIEL^*u^*v^*$ و $CIEL^*a^*b^*$ تبدیل می‌کنیم. بردار رنگی $\{L^*,a^*,b^*,v^*,u^*\}$ برای هر تصویر با استفاده از معادله‌هایی که در [17] ارجاع داده شده است محاسبه می‌شود. در این‌جا فقط پارامترهای $\{a^*, u^*\}$ استفاده شده است چون تفاوت اصلی بین لب و ناحیه چهره رنگ قرمز لب می‌باشد و در دو بردار انتخاب شده این رنگ مؤثرتر است.

۲ – ۶ – ۳ قطعه بندی [1]

بعد از پیش پردازش و تبدیل تصویر به فضای رنگی ذکر شده در بالا فرآیند قطعه‌بندی به صورت زیر انجام می‌شود.

۱. دو مؤلفه برداری $\{a^*, u^*\}$ به یکدیگر اضافه شده و اندازه تصویر برای تطابق با تصویر اصلی تغییر می‌کند.

۲. تبدیل هار صورت گرفته و ضرب انجام شده است و چهار ماتریس مختلف (dA, dH, dV, dD) مشخص می‌شوند.

ماتریس dA برای استخراج ناحیه لب کافی می‌باشد. پس سه ماتریس دیگر در نظر گرفته نشده است و در نهایت فیلترینگ شکل شناسی[2] و پس پردازش برای افزایش دقت بکار برده شده است. شکل لب

[1] Segmentation
[2] Morphological Filtering

با این روش استخراج شده و ویژگی‌هایی همچون پهنا و ارتفاع و زوایای گوشه لب و میانگین فاصله عمودی بین نقاط محاسبه می‌شود.

۲ – ۷ آنالیز مؤلفه‌های خاص

عملکرد این روش به این صورت است که ابتدا میانگین داده‌ها (بر روی هر بعد) را از داده‌ها کم می‌کند و داده‌های جدید با میانگین صفر تولید می‌نماید .سپس ماتریس کوواریانس داده‌های جدید محاسبه می‌شود .بردارهای ویژه یکه ماتریس کوواریانس را می‌توان به عنوان بردار ویژگی‌ها در نظر گرفت زیرا به نوعی پراکندگی داده‌ها را نشان می‌دهد .داده‌های نهایی، با ضرب بردارهای ویژگی در داده‌های با میانگین صفر به دست می‌آیند .

آنالیز مؤلفه‌های خاص برای فشرده‌سازی و استخراج ویژگی استفاده می‌شود. در [18] نمایش منیفلد[۱] بر اساس آنالیز مؤلفه‌های خاص برای شناسایی دیداری صحبت ارائه شده است. داده ویدیویی زمان واقعی توسط آنالیز مؤلفه‌های خاص فشرده شده و نقاط با بعد کم برای هر فریم که منیفلد را تعریف کند محاسبه می‌شوند. سیستم شناسایی شامل سه مرحله می‌باشد. در اولین گام تصویرها با استفاده از مؤلفه شبه رنگ[۲] از داده RGB لب‌ها استخراج می‌شوند [19] و با روش آستانه‌گیری[۳] مبتنی بر هیستوگرام[۴] یا سابقه‌نما لب‌ها قطعه‌بندی می‌شوند. دومین گام تولید ماکزیمم انتظار[۵] منیفلدها و انجام درون‌یابی و نمونه‌گیری دوباره از منیفلدها می‌باشد. سومین گام کلاسه‌بندی منیفلدها است.

[1] Manifold
[2] Pseudo-hue
[3] Thresholding
[4] Histogram
[5] Expectation Maximization

۲- ۷ – ۱ زمینه ریاضی EM – PCA

ماکزیمم انتظار آنالیز مؤلفه‌های خاص تعمیمی از روش آنالیز مؤلفه‌های خاص به وسیله ترکیب مزایای الگوریتم Em در بخش‌های تخمین مقادیر ماکزیمم احتمال برای اطلاعات از دست رفته می‌باشد. این روش در دو مرحله مجزا صورت می‌گیرد.

مرحله E : $W = (V^T V)^{-1} V^{-1} A$

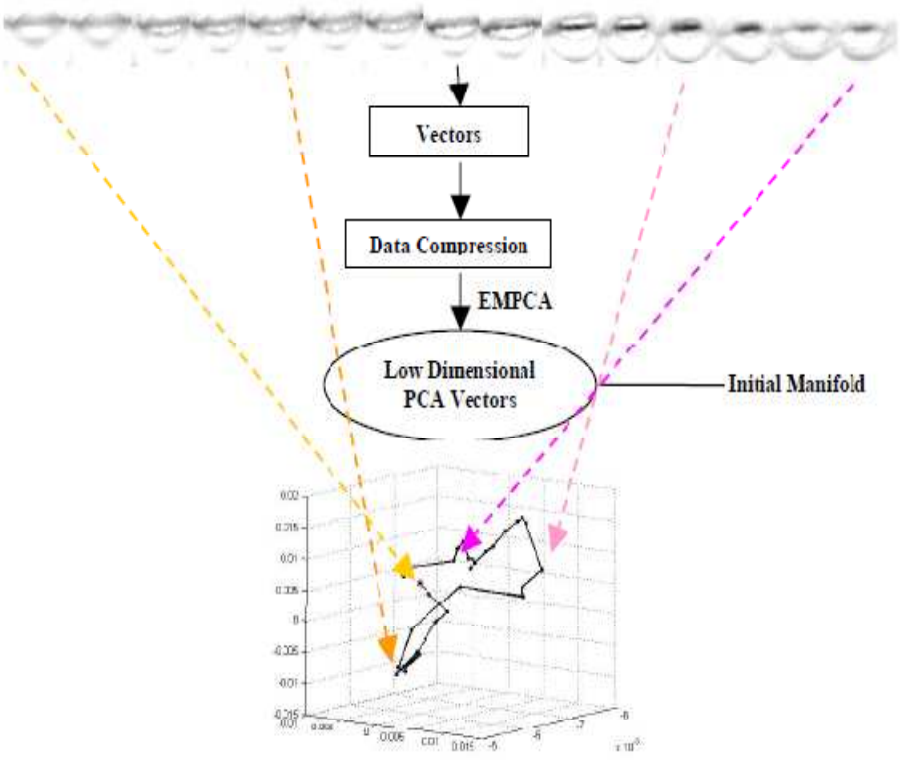
مرحله M : $V_{new} = A W^T (W W^T)^{-1}$

W ماتریس حالت‌های مجهول و V بردار داده و A داده مشاهده است.

۲- ۷- ۲ تولید منیفلد از تصویر ورودی

لب‌ها در هر فریم قطعه‌بندی شدند و داده سطح خاکستری اطراف ناحیه لب استخراج می‌شود و برای تولید فضایی با بعد کم توسط فرآیند EM – PCA استفاده می‌شود. سپس، داده سطح خاکستری روی این فضا طرح‌ریزی[1] شده و برای هر فریم یک نقطه یا بردار با بعد کم محاسبه می‌شود. نقاط ویژگی به دست آمده بعد از طرح‌ریزی داده روی فضای EM-PCA با بعد کم، با مرتب‌سازی فریم‌ها به صورت افزایشی نسبت به زمان توسط چندین خط به هم متصل می‌شوند. در شکل (۲- ۶) می‌بینیم.

[1] Projected

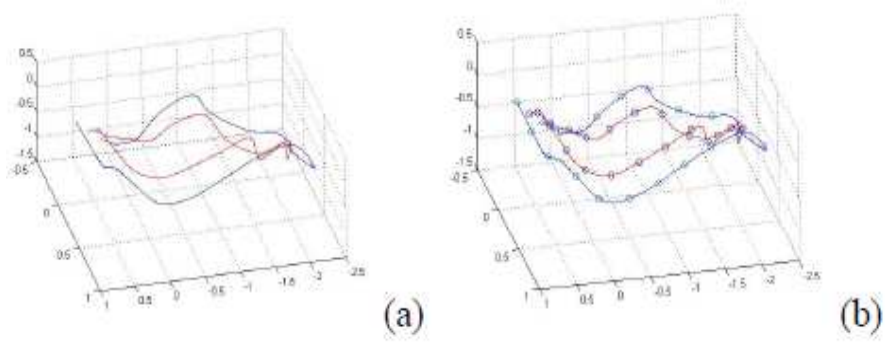


شکل ۲- ۶ فرآیند تولید منیفلد

منیفلدها مستقیماً برای شناسایی دنباله تصویر نمی‌توانند استفاده شوند به این دلیل که تعداد فریم-های موجود در دنباله ورودی متغیر است و وابسته به پیچیدگی کلمه‌ی بیان شده است. برای حل این مشکل نیاز به درون‌یابی منیفلدها برای دستیابی به یک سطح پیوسته و نمونه‌گیری مجدد آن به طور یکنواخت با استفاده از تعدادی نقاط کلیدی از پیش تعیین شده داریم.

برای درون‌یابی از نوار باریک مکعبی[1] استفاده شده است. که استفاده از آن باعث ایجاد سطح یکنواخت برای منیفلد و کاهش اثر نویز می‌شود. در شکل (۲- ۷) نمونه درون‌یابی شده (a) و نمونه‌گیری شده (b) نشان داده شده است.

[1] Cubic Spline



شکل ۲- ۷ (a) نتیجه درون‌یابی منیفلد (b) نمونه‌گیری دوباره از منیفلد درون‌یابی شده با ۲۰ نقطه کلیدی

۲- ۸ تبدیل کسینوسی گسسته

تبدیل کسینوسی از روش‌های مبتنی بر تصویر است که علاوه بر استخراج ویژگی از تصاویر، برای فشرده‌سازی تصویر نیز کاربرد دارد. در ویدیو، برای حرکت لب این تبدیل ساختاری سه بعدی دارد. با فرض ویژگی‌های حرکتی[1] لب از تبدیل کسینوسی گسسته (DCT) سه بعدی برای استخراج ویژگی استفاده شده است[20] .

۲- ۸- ۱ مدل‌سازی بر اساس 3-D DCT

برای محاسبه DCT سه بعدی می‌توانیم از ترکیب سه DCT یک بعدی استفاده کنیم.

$$X(l,m,n) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} x(i,j,k).\, C_{li}.\, C_{mj}.\, C_{nk}$$

رابطه (۲- ۳۳)

که x(i ,j ,k) مقدار پیکسل واقع شده در مختصات i, j, k را در ویدیو نشان می‌دهد.

$$C_{li} .C_{mj} . C_{nk} = \cos[\frac{\pi}{N}(i + 1/2)l] \cos[\frac{\pi}{N}(j + 1/2)m] \cos[\frac{\pi}{N}(k + 1/2)n]$$

رابطه (۲- ۳۴)

[1] Motion

از ساختار مدل مخفی مارکوف سه بعدی استفاده شده است. احتمال گذر حالت‌ها و سایر احتمال‌ها محاسبه شده و از الگوریتم ویتربی[1] برای شناسایی استفاده شده است.

کار بر روی پایگاه داده VidTIMIT انجام گرفته و ۱۸ نفر برای آموزش و ۵ نفر برای آزمایش انتخاب شدند. ۳۱ ضریب تبدیل کسینوسی گسسته که $l+ m + n \leq 3$ را برآورده می‌کند را، از مکعب‌های ۸×۸×۸، به عنوان بردار ویژگی در نظر گرفته‌اند. در [21] از اطلاعات حرکت لب برای شناسایی صحبت استفاده شده که در ادامه بیان شده است.

۲- ۸- ۱-۱ استخراج ویژگی حرکت لب

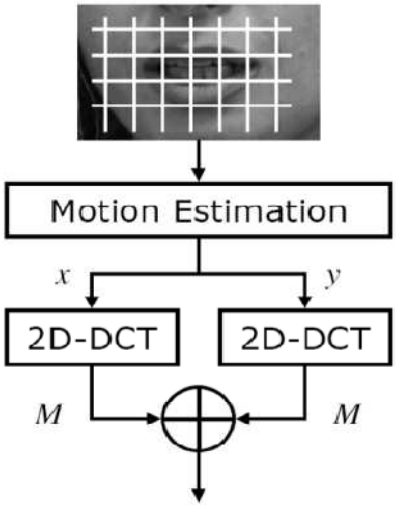
برای این منظور مراحلی چون پیش پردازش، تخمین حرکت لب، جداسازی زمانی، جداساز بیزین[2] انجام می‌شود. دو روش استخراج ویژگی حرکت بر اساس شبکه[3] و بر اساس کانتور بیان شده است.

۲- ۸- ۱- ۲ استخراج ویژگی حرکت مبتنی بر شبکه

شبکه‌ای به اندازه $Gx \times G_y$ روی ناحیه لب استخراج شده از تصویر در نظر گرفته می‌شود. برای تخمین حرکت لب از تطبیق بلوکی سلسله مراتبی استفاده شده است. فرآیند تخمین حرکت، ماتریس‌های دو بعدی V_x, V_y که شامل مؤلفه‌های x, y از بردارهای حرکت در نقاط شبکه است را ایجاد می‌کند. از این ماتریس‌ها به صورت مجزا تبدیل کسینوسی گسسته دو بعدی گرفته می‌شود. M ضریب اول DCT در طول مرحله پویش شکسته[4] یا همان اسکن زیگزاگ، در دو جهت x, y برای تشکیل بردار ویژگی f از بعد 2M ترکیب می‌شوند. این بردار ویژگی حرکت شبکه متراکم را نمایش می‌دهد و به عنوان f_{GRD} معرفی می‌شود. در شکل (۲- ۸) نشان داده شده است. این تبدیل دو فایده

[1] Viterbi
[2] Bayesian Discriminative
[3] Grid
[4] Zig-Zag Scan

دارد. اولین فایده این است که بعد ویژگی‌ها را با حذف مؤلفه‌های فرکانس بالا از سیگنال حرکت کاهش می‌دهد که این مؤلفه‌های به خاطر نویز ایجاد می‌شوند. دومین فایده این است که DCT بردار ویژگی را ناهمبسته می‌سازد.

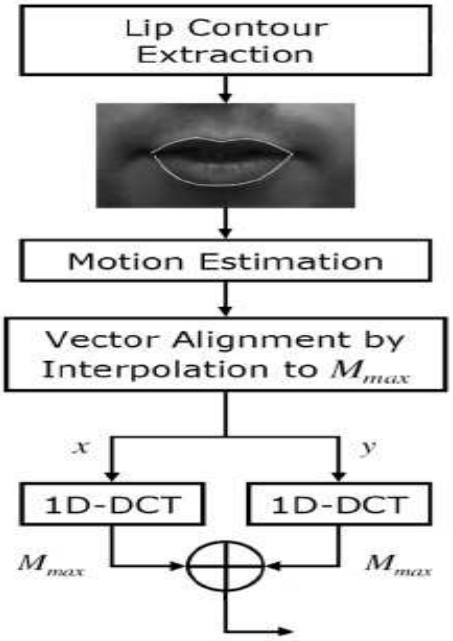


شکل ۲- ۸ نمودار بلوکی برای استخراج ویژگی‌های حرکت مبتنی بر شبکه

$f_{GRD} = \{ f_x^1 , f_y^1 , f_x^2 , f_y^2 , \ldots\ldots , f_x^M , f_y^M \}$

۲- ۸- ۱ - ۳ استخراج ویژگی حرکت مبتنی بر کانتور

در این روش بردارهای حرکت روی پیکسل‌های کانتور لب محاسبه می‌شوند. دو دنباله از مؤلفه‌های حرکت x , y روی کانتور به صورت جداگانه با DCT یک بعدی تبدیل می‌شوند. طول دنباله‌ی نتیجه در هر جهت، از یک فریم به دیگری مطابق با تغییر شکل لب ممکن است تغییر کند. برای دستیابی به بردار ویژگی با اندازه ثابت قبل از تبدیل، طول دنباله به مقدار ثابتی توسط درون‌یاب خطی نرمالیزه می‌شود. این مقدار M_{max}، ماکزیمم تعداد نقاط کانتور به دست آمده در هر فریم لب از دنباله موجود می‌باشد. ضرایب DCT به صورت مجزا برای x , y محاسبه می‌شود و در نهایت برای تشکیل بردار ویژگی که f_{CTR} تعریف می‌شود به یکدیگر الحاق می‌شوند. شکل (۲- ۹) این فرآیند را نشان می‌دهد.



شکل ۲- ۹ استخراج ویژگی حرکت مبتنی بر کانتور

$F_{CRT} = \{ f_x^1 , f_y^1 , f_x^2 , f_y^2 ,\ldots\ldots, f_x^{Mmax} , f_y^{Mmax}\}$

۲-۸-۲ استخراج ویژگی از ناحیه مورد نظر

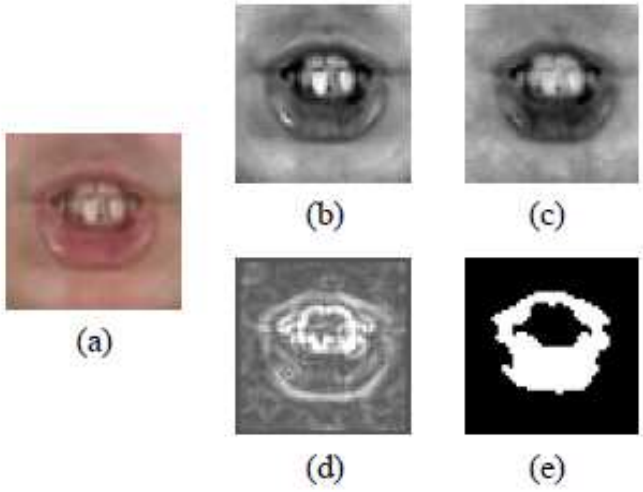
در [22] ابتدا از مجموعه تصاویر صورت آشکار شده و سپس ناحیه شامل دهان استخراج شده و ویژگی از این ناحیه به دست آمده است. بعد از اینکه ناحیه مورد نظر[1] استخراج و سایز تصاویر به ۴۸*۴۸ تغییر داده شده و اثر ۴ ناحیه مختلف بر روی دقت شناسایی بررسی شده است.

این نواحی توسط چهار پردازش مختلف روی تصویر به دست آمده‌اند. ناحیه اول که تصویر سطح خاکستری نرمالیزه شده از تصویر اصلی می‌باشد و ناحیه دوم از اعمال تبدیل Fisher [2] به تصویر اصلی

[1] Region Of Interest
[2] Fisher

و ناحیه سوم و چهارم به ترتیب از اعمال آشکار ساز لبه سوبل[1] به ناحیه دوم و باینری کردن[2] ناحیه دوم حاصل شده‌اند. که ناحیه اول توسط رابطه زیر نرمالیزه شده است.

$$x = \frac{x-\mu}{6\sigma} + 0.5$$ رابطه (۲-۳۵)



شکل ۲-۱۰ تصویر اصلی و چهار ناحیه پردازش شده برای استخراج ویژگی

۲-۸-۲-۱ استخراج ویژگی‌های دیداری

تبدیل کسینوسی گسسته برای محاسبه بردار ویژگی از این نواحی صورت گرفته است که دلیل استفاده از این روش به خاطر فشرده‌سازی زیاد انرژی سیگنال ورودی بر روی تعداد کمی از ضرایب و نیز قابلیت پیاده‌سازی سریع این تبدیل می‌باشد. بعد از گرفتن تبدیل کسینوسی با اسکن زیگزاگ ماتریس ضرایب به برداری تبدیل که با انتخاب چند ضریب اول از این بردار که بیشینه انرژی تصویر را نشان می‌دهند بردار ویژگی استخراج شده است. شناسایی توسط CHMM که مجموعه‌ای با ۳ حالت و ۳ ترکیب گوسی بر حالت می‌باشد انجام شده است. در این کار از ۱۰ گوینده که شامل ۸ مرد و ۲ زن می‌باشد که ۸۱ کلمه چینی را ۴ مرتبه تکرار کرده‌اند و با نرخ ۲۵ فریم بر ثانیه ضبط شده و سایز

[1] Sobel
[2] Binarization

تصاویر ۲۴۰*۳۲۰ می‌باشد استفاده شده است. ابعاد ضرایب کسینوسی از ۲۹ تا ۱۲۹ با گام۱۰ تغییر داده شده و هر بار به ازای تعداد مشخصی از ضرایب و برای ناحیه‌ای از ۴ ناحیه ذکر شده در بالا، دقت شناسایی محاسبه شد.

۲-۸-۳ تبدیل کسینوسی و LSDA[1]

در [23] یک روش جدید برای استخراج ویژگی برای لب‌خوانی ارائه شده است. تبدیل کسینوسی همراه با LSDA بکار گرفته شده و با دو روش دیگر DCT + PCA , DCT + LDA مقایسه شده است.

۲-۸-۳-۱ پیش پردازش

قبل از ورود به مرحله استخراج ویژگی ابتدا باید ویدیو به بخش‌های کلمه تقسیم‌بندی و سپس ناحیه مورد نظر از فریم‌های ویدیو گرفته شود.

۲-۸-۳-۲ روش DCT

بعد استخراج ناحیه دهان از تصاویر صورت، سایز تصاویر به ۶۴ * ۴۸ تغییر داده و تبدیل DCT گرفته شده و ضرایب گوشه چپ و بالای ماتریس به عنوان ضرایب مهم کسینوسی انتخاب شده‌اند.

۲-۸-۳-۳ DCT + PCA

آنالیز مؤلفه‌های خاص یک روش غیر نظارتی است که میانگین مربع خطا را می نیمم می‌کند که از این حیث تبدیلی بهینه است. در این روش بعد از اعمال تبدیل کسینوسی به تصویر و انتخاب ضرایب

[1] Locality Sensitive Discriminant Analysis

کسینوسی مهم آن‌ها را به عنوان ورودی به PCA داده تا کاهش بعد صورت گیرد. که عملکرد این دو روش به همراه هم بهتر از تبدیل کسینوسی به تنهایی می‌باشد.

DCT + LDA ۴-۳-۸-۲

آنالیز مجزاساز خطی روشی بر اساس ماتریس‌های پراکندگی[1] درون کلاس‌ها S_w و ماتریس پراکندگی بین کلاس‌ها S_b می‌باشد. که به یافتن ماتریس تبدیلی که ماتریس پراکندگی بین کلاس‌ها را ماکزیمم و ماتریس پراکندگی درون کلاس‌ها را می‌نیمم می‌کند کمک می‌کند.

$$a_{opt} = argmax\ \frac{a^T S_b\, a}{a^T S_w\, a}$$ رابطه (۲-۳۶)

$$S_w = \sum_{i=1}^{c} \sum_{x \epsilon D_i} (x - m_i)(x - m_i)'$$ رابطه (۲-۳۷)

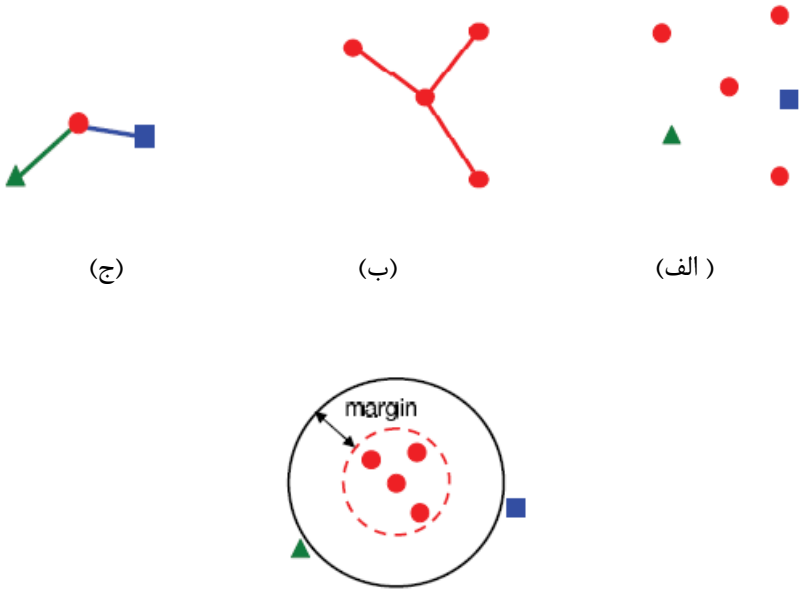
$$S_b = \sum_{i=1}^{c} l_i (m_i - m)(m_i - m)'$$ رابطه (۲-۳۸)

جایی‌که m بردار میانگین کل نمونه‌ها و m_i بردار میانگین کلاس iام و l_i تعداد کلاس iام و c تعداد کلاس‌ها است. ضرایب مهم کسینوسی به LDA اعمال و بردار ویژگی از آن محاسبه شد.

DCT + LSDA ۵-۳-۸-۲

این روش ساختار هندسی و تفکیک‌کنندگی را با هم در نظر می‌گیرد که برای این کار دو گراف ، گراف بین کلاس‌ها G_b و گراف درون کلاس‌ها G_w در نظر گرفته می‌شود. مجموعه‌ای شامل نقاط همسایه با x_i که برچسب یکسانی دارند را با $N_w(x_i)$ و نقاطی که برچسب آن‌ها متفاوت است یا به عبارتی مربوط به کلاس‌های مختلفند با $N_b(x_i)$ نشان داده می‌شود. $y = (y_1 , y_2 , \ldots\ldots , y_m)^T$ مدلی است که گراف بین کلاس و گراف درون کلاس را به یک خط نگاشت می‌کند به طوری‌که نقاط اتصال از G_w تا جایی که ممکن است نزدیک به هم و نقاط اتصال G_b از هم دور بمانند.

[1] Scatter Matrix



(ج) (ب) (الف)

(د)

شکل ۲-۱۱ (الف) نقاط با رنگ و شکل مشابه در یک کلاس قرار می گیرند. (ب) گراف درون کلاس نقاط با برچسب یکسان را متصل می کند. (ج) گراف بین کلاس نقاط با بر چسب متفاوت را متصل می کند. (د) بعد از اعمال LSDA فاصله بین کلاس های متفاوت ماکزیمم شده است.

معیار برای انتخاب این نقشه یا مدل بهینه سازی دو تابع زیر است.

$$\min \sum_{i,j}(y_i - y_j)^2 W_{w,ij}$$ رابطه (۲-۳۹)

$$\max \sum_{i,j}(y_i - y_j)^2 W_{b,ij}$$ رابطه (۲-۴۰)

جایی‌که W_b , W_w ماتریس‌های وزن گراف‌ها می‌باشند و داریم:

$$W_{b,ij} = \begin{cases} 1, & \text{if } x_i \in N_b(x_j) \text{ or } x_j \in N_b(x_i) \\ 0, & \text{در غیر اینصورت} \end{cases}$$ رابطه (۲-۴۱)

$$W_{w,ij} = \begin{cases} 1, & \text{if } x_i \in N_w(x_j) \text{ or } x_j \in N_w(x_i) \\ 0, & \text{در غیر اینصورت} \end{cases}$$ رابطه (۲-۴۲)

جاییکه $N_b(x_i)$, $N_w(x_i)$ برای k همسایه نزدیک بین کلاس و درون کلاس قرار می‌گیرند. ($N_w(x_i)$ همسایه‌هایی که برچسب یکسان با x_i دارند و $N_b(x_i)$ همسایه‌هایی که برچسب متفاوت دارند را شامل می‌شود. بعد از یافتن بردار طرح [1] داریم $y^T= a^T X$.

تابع هدف برای رابطه (۲-۳۹)، با باز کردن رابطه به $\max a^T X W_w X^T a$ و برای رابطه (۲-۴۰) به $\max a^T X L_b X^T a$ کاهش داده می‌شود. که $L_b= D_b - W_b$ لاپلاسین[2] ماتریس G_b است. که D_b ماتریسی قطری است که ورودی‌هایش مجموع ستون‌ها یا سطرهای W_b می‌باشد.

$D_{b,ii} = \sum_j W_{b,ij}$, $D_{W,ii} = \sum_j W_{W,ij}$ رابطه (۲-۴۳)

در نهایت مسئله بهینه‌سازی به یافتن $\arg\max a^T X(\alpha L_b+(1-\alpha)W_w)X^T a$ کاهش می‌یابد با توجه به اینکه $a^T X D_w X^T a=1$ یا $y^T D_w y=1$ و α مقدار ثابتی که $0 \le \alpha \le 1$ است.

با حل رابطه $X(\alpha L_b+(1-\alpha)W_w)X^T a = \lambda X D_w X^T a$ بردار ستونی شامل $a_1 , a_2 , \ldots\ldots , a_d$ به دست می‌آید.

مانند مراحل قبل پس از اعمال تبدیل کسینوسی به ناحیه مورد نظر و استخراج ضرایب مهم ، آن‌ها به LSDA داده شده و خروجی به عنوان بردار ویژگی در نظر گرفته شده است.

این روش‌ها بر روی پایگاه داده (HIT Bi CAVDB) [3] که شامل ۱۰۰۰ کلمه است که هر کدام ۳ مرتبه تکرار شده و فایل‌ها دارای فرمت ’Avi‘ هستند و به‌صورت دستی به فریم‌هایشان مطابق با سیگنال‌های صوتی سگمنت‌بندی شده‌اند اعمال شده است. که ۹۶ سیلاب متفاوت چینی (کلاس) را شامل می‌شود و با نرخ ۲۵ فریم بر ثانیه ضبط و سایز تصاویر ۲۵۶*۲۵۶ می‌باشد.

[1] Projection Vector
[2] Laplacian
[3] Harbin Institute of Technology Bimodal Chinese Audio-Video Database

۲-۸-۳-۶ ماتریس انتقال ویژگی

چون برای یک کلمه ، تعداد متفاوتی فریم برای نمونه‌های مختلف وجود دارد بنابراین غیر ممکن است که برای آموزش ماتریس انتقال استفاده شوند. پس شکل لب به ده کلاس مطابق با ارتفاع و پهنای لب و گردشدگی و دندان‌ها دسته‌بندی می‌شود. از هر نوع ۶۰ نمونه برای آموزش ماتریس انتقال ویژگی LSDA انتخاب شده است. در نهایت از این ماتریس برای استخراج ویژگی نهایی استفاده و چون شکل لب ها به ۱۰ کلاس دسته‌بندی شده ، بعد ویژگی ۹ در نظر گرفته شده و برای یک کلمه شامل n فریم برداری به سایز n*9 به دست آمده است. برای شناسایی DTW[1] بکار گرفته شده و روش‌ها با هم مقایسه شده‌اند که نتایج حاصل از روش DCT + LSDA از سایر روش‌ها بهتر بوده است.

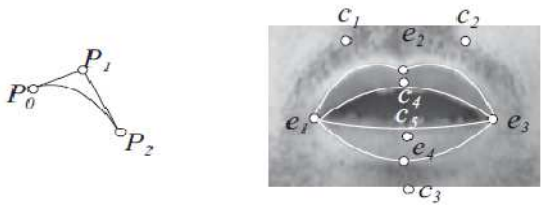
۲-۹ مدل لب با منحنی بیزیر[2]

در [24] یک مدل لب جدید مبتنی بر منحنی‌های بیزیر برای محاسبه حرکت‌های لب استفاده شده است. این مدل توسط تعدادی نقطه که به وسیله مدل شکل فعال شکل گرفته‌اند تعریف می‌شود. در این‌جا ابتدا صورت و لب‌ها آشکار می‌شوند. بعد لب‌ها توسط پنج منحنی بیزیر مدل می‌شوند. که هر کدام توسط دو نقطه انتهایی p_0 , p_2 و یک نقطه کنترل p_1 مانند شکل (۲- ۱۱) تعریف و به صورت زیر نوشته می‌شوند.

$$P(t)= \phi_0(t)\, p_0 + \phi_1(t)\, p_1 + \phi_2(t) p_2 \qquad \text{رابطه (۲- ۴۴)}$$

$$\phi_0(t)= (1-t)^3 , \ \phi_1(t)= 3t(1-t)^2 \ , \ \phi_2(t) = (3t^2 - 2t^3) \ , \qquad t\epsilon\, [0,1]$$

[1] Dynamic Time Warping
[2] Bezier Curves



شکل ۲- ۱۲ سمت چپ منحنی بیزیر و سمت راست مدل لب

مدل شامل چهار نقطه انتهایی e_4 ,e_3 ,e_2 ,e_1 و پنج نقطه کنترل c_5 , c_4 , c_3 , c_2 , c_1 می‌باشد. این مدل ۱۵۰ نقطه ویژگی (جایی که هر منحنی شامل ۳۰ نقطه است) را که مرزهای لب را تعریف می-کنند فشرده می‌کند. مدل هر شکل از ویزم‌های آلمانی را تطبیق می‌دهد و قادر است که حرکت‌های لب را محاسبه کند. که حرکت‌های لب توسط مدل شکل فعال می‌تواند توصیف شود.

در [25] سیستم دیداری انسان (HVS) مبتنی بر معیارهای کیفیت تصویر به ویژه شباهت ساختاری موجک پیچیده[۱] (CW-SSIM) و درستی اطلاعات تصویری[۲] (VIF) به عنوان معیارهای تشابه استفاده شده است.

CW-SSIM برای هر باند فرعی از اولین تجزیه موجک محاسبه می‌شود و سپس، میانگین این مقادیر چندین معیار CW-SSIM برای هر تصویر به دست می‌دهد. که جزئیات آن در [26] بیان شده است.

فرهنگی[۳] از مصوت‌ها که شامل ۴ ویدیوی ضبط شده برای هر مصوت است جمع‌آوری شده و با ویدیوی آزمایش مقایسه می‌شود. SSIM تابعی از روشنایی، اختلاف روشنایی[۴] و تابع ساختار تصویر است.

$$S(x , y) = I(x , y).\ c(x , y).s(x , y)$$

$$= \left(\frac{2\mu_x\,\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1}\right).\left(\frac{2\sigma_x\,\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}\right).\left(\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}\right) \qquad \text{رابطه (۲- ۴۵)}$$

[1] Complex Wavelet Structural Similarity
[2] Visual Information Fidelity
[3] Dictionary
[4] Contrast

چون SSIM عملکرد خوبی نداشته از CW-SSIM استفاده شده است. که در زیر روابط آن ذکر شده است.

$\tilde{S}(c_x , c_y) = \tilde{m}(c_x , c_y) . \tilde{p}(c_x , c_y)$

$$= \frac{2\sum_{i=1}^{N}|c_{x,i}||c_{y,i}|+k}{\sum_{i=1}^{N}|c_{x,i}|^2+\sum_{i=1}^{N}|c_{y,i}|^2+k} * \frac{2|\sum_{i=1}^{N}c_{x,i}c_{y,i}^*|+k}{2\sum_{i=1}^{N}|c_{x,i}c_{y,i}^*|+k}$$ رابطه (۲- ۴۶)

$c_x = \{c_{x,i} | i= 1,2,\ldots,N\}$, $c_y = \{c_{y,i} | i= 1,2,\ldots,N\}$

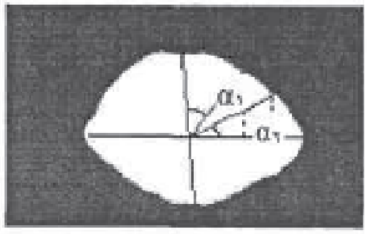
که c_x و c_y ضرایب موجک هستند. k مقدار ثابتی برای پایداری است.

همه تصاویر ویدیویی ابتدا به فریم‌هایشان با نرخ ۲۵ فریم بر ثانیه شکسته می‌شوند. هر ویدیو حدود ۱۰۰ فریم دارد. چند فریم با فریم‌های متناظرش در دیگر دنباله‌ها توسط CW-SSIM و VIF مقایسه می‌شود.

۲-۱۰ جداسازی ناحیه لب با کا- مینز[1]

در [27] از روشی تلفیقی از روش‌های استخراج رنگ قرمز، روش کا- مینز و باینری کردن تصاویر برای استخراج ناحیه دهان در فریم‌ها استفاده شده است. در این مطالعه علاوه بر ویژگی‌های ارتفاع و پهنای دهان، زاویه گشودگی عمودی و افقی دهان که در شکل (۲- ۱۲) نشان داده شده است نیز استخراج می‌شود.

[1] K- means



شکل ۲- ۱۳ زاویه گشودگی افقی α_2 و زاویه گشودگی عمودی α_1

بعد از تعیین ناحیه دهان بر روی رشته تصاویر رنگی مربوط به کلمات دو سیلابی فارسی، سیلاب‌ها جداسازی شده و مصوت موجود در هر یک از سیلاب‌ها شناسایی می‌شود.

در [28] توسط قطعه‌بندی و روش‌های مدل‌سازی یک بردار ویژگی تصویری متشکل از ویژگی‌های داخلی و خارجی دهان از دنباله تصویر لب برای شناسایی به دست آمده است. از نمایش نوار باریک[1] برای تبدیل ویژگی‌های نمونه‌گیری شده زمان گسسته از فریم‌های ویدیویی به حوزه پیوسته استفاده شده است.

بعد از ایجاد مدل‌های مناسب کلمه از ضرایب spline، روش کلاسه‌بندی ماکزیمم احتمال (EM) برای شناسایی اتخاذ شده است. از مدل شکل فعال استفاده شده، پهنا و ارتفاع لب به دست آورده شده و نرمالیزه شده و همچنین، بردارهای ویژه محاسبه و سه مقدار اول وزن‌ها انتخاب شده‌اند. از ویژگی‌های داخلی دهان نیز مساحت ناحیه دندان‌ها و گشودگی داخلی دهان که نرمالیزه شده هستند نیز استفاده شده است.

[1] Spline

فصل سوم : روش های استخراج ناحیه دهان و سیستم‌های تشخیص

۳-۱ مقدمه

طبق آن‌چه که در قسمت‌های قبل بیان کردیم روش‌های مختلفی برای استخراج ویژگی وجود دارد. ویژگی‌هایی چون پهنای دهان، ارتفاع دهان، ارتفاع و پهنای لب بالایی و لب پایینی، گشودگی افقی و یا عمودی دهان، زوایه‌های گشودگی عمودی و افقی، زاویه بین نقاط گوشه چپ و راست لب، فاصله عمودی نقاط روی مرز لب، ویژگی‌های حرکت لب، ضرایب تبدیل کسینوسی گسسته، هیستوگرام، فواصل شعاعی یا همان فاصله نقاط روی مرز از مرکز دهان و همچنین، ویژگی‌هایی که همیشه قابل رؤیت نیستند همچون ارتفاع زبان زیر لب بالایی، ارتفاع زبان بالای لب پایینی، ارتفاع زبان بین دندان-ها، ارتفاع دندان‌های بالایی و پایینی را می‌توان نام برد.

دقت استخراج ویژگی‌های لب برای شناسایی مهم می‌باشد. یکی از عمومی‌ترین روش‌ها برای استخراج ویژگی استفاده از مقادیر سطح خاکستری و آشکارسازی لبه است. اما این روش‌ها به دلایلی همچون وجود سبیل و ریش ضعیف کار می‌کنند. روش دیگر استفاده از لبه‌های افقی است که توسط کانوال[1] تصویر با عملگر لبه D_y می‌باشد چون ناحیه دهان در جهت افقی مقدار لبه بزرگتری دارد. سپس تصویر نتیجه آستانه‌گذاری می‌شود که این روش نیز نتایج قابل قبولی نخواهد داشت. به همین دلیل به سراغ طیف رنگی رفته و با مولفه‌های رنگی به شناسایی ناحیه لب پرداخته شده است.

$$DY = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix}$$ رابطه (۳-۱)

[1] Convolving

۳-۲ آشکارسازی ناحیه لب

۳-۲-۱ آنالیز ترکیب رنگ لب وپوست

در فضای رنگی[1] RGB ، پیکسل‌های پوست و لب مؤلفه‌های کاملاً متفاوتی دارند. برای هردو مؤلفه قرمز یکسان است، در ترکیب رنگ پوست مؤلفه سبز نسبت به آبی بزرگتر است و برای لب‌ها این دو مؤلفه تقریباً یکسان می‌باشد. اختلاف بین مؤلفه قرمز و سبز برای لب‌ها نسبت به پوست بزرگتر می-باشد. در [29] یک تعریف شبه رنگی بیان شده که این اختلاف را نشان می‌دهد و به صورت زیر محاسبه می‌شود.

$$h(x,y) = \frac{R(x,y)}{R(x,y)+G(x,y)}$$ رابطه (۳-۲)

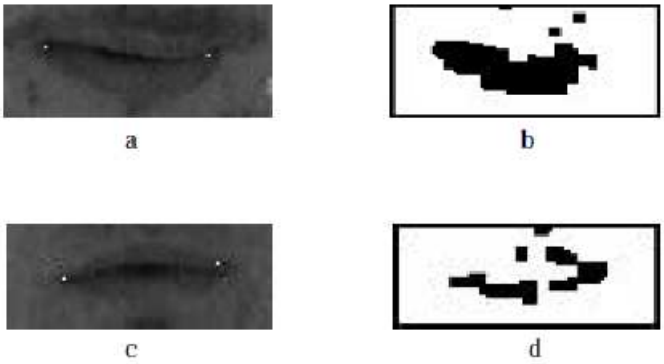
R ,G به ترتیب مؤلفه‌های سبز و قرمز هستند.

در [30] از معادله زیر برای شناسایی پیکسل‌های لب استفاده شده است.

$$L_{lim} \leq \frac{R}{G} \leq U_{lim}$$ رابطه (۳-۳)

L_{lim} , U_{lim} بیشترین و کمترین آستانه‌ای هستند که مقدار مؤلفه قرمز به سبز از پیکسل‌های لب را تعریف می‌کنند. بعد از حذف برخی از پیکسل‌های نادرست و انجام عملیات شکلی[2] (گشودن[3] و بستن[4]) تصویر نتیجه به صورت زیر به دست آمده است. در این روش گوشه‌های افقی لب شناسایی شده است.

[1] Red , Green , Blue
[2] Morphologically
[3] Opening
[4] Closing



شکل ۳-۱ نتیجه حاصل از آنالیز ترکیب رنگ پوست و لب و نقاط گوشه لب

۳-۲-۲ رنگ[1] و اشباع[2] و شدت روشنایی[3] (HSV)

فضای رنگی (HSV) روشنایی را از رنگ جدا می‌سازد بطوریکه تغییرات در روشنایی سبب تغییر زیادی در رنگ نخواهد شد. در [31],[32] مقادیر رنگ (Hue) برای محاسبه پیکسل‌های لب استفاده شده است. هردو الگوریتم مشابهی را برای محاسبه احتمال اینکه پیکسل مربوط به لب باشد استفاده کرده‌اند که به صورت زیر بیان شده است.

$$f(h)=\begin{cases}1-\dfrac{(h-h_o)^2}{w^2} & , \quad |h-h_o|\le w \\ 0 & , \quad otherwise\end{cases}$$

رابطه (۳-۴)

که این روش در شرایط ایده‌آل مناسب است و به خوبی برای تصاویر متفاوت و در شرایط مختلف عمل نمی‌کند.

[1] Hue
[2] Saturation
[3] Value

۳-۲-۳ حذف مؤلفه قرمز

این روش از مقادیر رنگ سبز و آبی استفاده و از نسبت مؤلفه سبز به آبی، ناحیه لب یافت شده است. برای این کار ابتدا تصویر با یک فیلتر گوسی[1] برای حذف نویز کانوال و سپس لگاریتم نسبت رنگ سبز سبز به آبی محاسبه می‌شود.

$$\log\left(\frac{G}{B}\right) \leq \beta \qquad \text{رابطه (۳- ۵)}$$

در واقع برقرار بودن نامساوی رابطه (۳- ۵) تعلق یا عدم تعلق هر پیکسل از تصویر را به لب تعیین می‌کند. در [33] از این روش استفاده شده است که مقدار $\beta = (\mu - 1.05*\sigma)$ تعریف شده است که σ انحراف استاندارد و μ میانگین مطابق با داده‌ی آماری هستند.

از تبدیلات رنگی دیگری نیز برای یافتن ناحیه لب استفاده شده است که در [33] , [29] به ترتیب از روابط زیر استفاده شده است.

$$h(x,y) = \frac{R(x,y)^3}{R(x,y)^3 + G(x,y)^3 + 1} \qquad \text{رابطه (۳-۶)}$$

$$\text{که } R_{cor}(x,y) = \frac{R(x,y)}{R(x,y) + (b-a) * L(x,y) + a} \qquad \text{رابطه (۳- ۷)}$$

$L(x,y)$ مقدار روشنایی است و $(a,b) = (0.4,0.8)$ در نظر گرفته شده است.

۳-۲-۴ الگوریتم کا- مینز

این الگوریتم در سال ۱۹۶۷ توسط مک کوئین[2] معرفی شد که داده‌ها را به k خوشه مجزا با مقدار متوسط C_j تقسیم‌بندی می کند و به صورت زیر بیان می‌شود:

[1] Gaussian Filter
[2] McQueen

k مرکز اولیه $C_1 , C_2 , ..., C_K$ برای k خوشه از میان داده‌های ورودی (x) بر اساس قاعده دلخواهی انتخاب می‌شود. که می تواند تصادفی یا بر اساس توزیع داده‌ها باشد.

در n امین مرحله معین می‌شود که هر داده متعلق به کدام خوشه است که بر اساس معیار نزدیک بودن داده به مرکز خوشه می‌باشد.

مقدار متوسط داده‌های اختصاص یافته به هر خوشه در مرحله n محاسبه شده و مقدار به دست آمده به عنوان مرکز جدید خوشه در مرحله n+1 در نظر گرفته می‌شود.

دو مرحله قبل آنقدر تکرار می‌شود تا دیگر محل همه مراکز خوشه‌ها نسبت به مرحله قبل تغییر چندانی نکند. در نهایت کیفیت خوشه‌بندی توسط تابع خطایی محاسبه می‌گردد. که مقدار این خطا برابر مجموع مربع خطای هر داده تا مرکز خوشه خود می‌باشد که میزان خطا به تعداد خوشه‌ها بستگی دارد. مهم‌ترین مسئله در این روش تعیین بهینه تعداد خوشه‌ها و مقدار اولیه مراکز خوشه‌ها می‌باشد. در [27] از این روش برای جداسازی ناحیه دهان استفاده شده است.

۳-۲-۴-۱ پیاده سازی الگوریتم

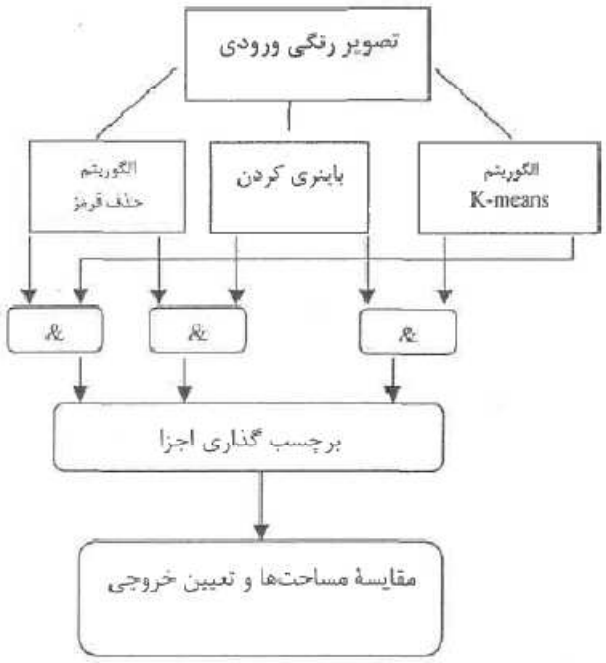
در این روش با توجه به اینکه هدف جداسازی ناحیه دهان از ناحیه غیر دهان می‌باشد تعداد خوشه‌ها برابر با ۲ و مقادیر اولیه‌ای برای مرکز این دو خوشه در نظر گرفته می‌شود. روند ناحیه‌بندی بدین صورت انجام گرفته که از پیکسل (۱و۱) تصویر شروع به بررسی شده و فاصله هر پیکسل تا دو مرکز انتخاب شده محاسبه می‌شود. اگر فضای رنگ انتخابی سطح خاکستری باشد داده به مرکزی تعلق دارد که فاصله سطح خاکستری آن پیکسل تا سطح خاکستری مرکز نسبت به سطح خاکستری مرکز دیگر، کمتر باشد. این روند برای تمامی پیکسل‌های تصویر انجام می‌گیرد و سپس با متوسط‌گیری روی سطح خاکستری پیکسل‌های تخصیص یافته روی هر خوشه، مراکز جدید دو خوشه محاسبه می-گردد و روند فوق آنقدر تکرار شده تا شرایط پایانی کا- مینز برقرار شود.

۳-۲-۵ شدت روشنایی و باینری کردن

یکی از ساده‌ترین روش‌ها برای جداسازی لب از سایر قسمت‌ها استفاده از آستانه‌ای مناسب است که با این روش می‌توان به تصاویری باینری شده دست یافت. در این صورت با تعیین مقدار آستانه مشخص می‌توان مقدار پیکسل‌هایی که شدت روشنایی آن‌ها از مقدار آستانه بیشتر است را برابر یک و آن‌هایی که شدت روشنایی آن ها از این آستانه کمتر است را برابر صفر قرار دهیم. البته باید توجه داشت که این روش به طور کامل قادر به جداسازی لب نخواهد بود.

۳-۲-۶ روش‌های ترکیبی

یکی دیگر از کارهایی که برای جداسازی لب از سایر قسمت‌های صورت می‌توان انجام داد استفاده از ادغام روش‌های بیان شده در قسمت‌های قبل است. در [27] روشی ترکیبی برای جداسازی ناحیه لب ارئه شده است. در این تحقیق از سه روش الگوریتم حذف قرمز و کا- مینز و روش باینری کردن استفاده شده و بعد برای ترکیب نتایج حاصل بین نواحی تشخیص داده شده توسط این الگوریتم‌ها دو به دو اجتماع گرفته شده و سپس نتایج حاصل از این عمل برچسب‌گذاری شده و در نهایت با مقایسه مساحت‌ها خروجی تعیین می‌شود.



شکل ۳-۲ الگوریتم جداسازی ناحیه لب

برای شناسایی دیداری از ویژگی‌هایی همچون ارتفاع و پهنای دهان و زاویه گشودگی افقی و عمودی دهان استفاده شده است. عملکرد هریک از این ویژگی‌ها به صورت مستقل و ترکیبی بررسی شده و مشخص شده که استفاده همزمان از مؤلفه زاویه گشودگی افقی دهان به همراه ارتفاع و پهنای دهان بهترین امکان جداسازی بین سیلاب‌ها را ایجاد کرده و بنابراین بعد از جداسازی سیلاب‌ها به تشخیص مصوت پرداخته شده است. شبکه عصبی دو لایه با ۲۵ نرون میانی و ۶ نرون خروجی متناظر با تعداد کلاس‌ها به کار گرفته شده و برای آموزش از روش RPROP[1] استفاده شده است. یک بار آموزش و تست فقط بر روی زاویه‌های گشودگی افقی و بار دوم همزمان بر روی زاویه‌های گشودگی افقی و عمودی انجام شده و پس از انجام این مرحله و وزن‌دهی خروجی‌ها و تعیین ماکزیمم به عنوان خروجی، مصوت موجود در سیلاب تعیین گردیده است.

[1] Resilient Propagation

استفاده از هر یک از این روش‌ها و استخراج ناحیه لب باعث کاهش ابعاد و پیچیدگی‌ها و عملکرد بهتر ویژگی‌ها و در نتیجه افزایش دقت شناسایی خواهد شد بنابراین بعد از استخراج این ناحیه بهتر است ویژگی‌های مد نظر از آن استخراج شوند.

۳-۳ روش‌های کلاسه‌بندی و شناسایی

برای شناسایی دیداری صحبت روش‌های مختلفی همچون مدل مخفی مارکوف (HMM)، شبکه‌های عصبی (NN)[1] و نزدیک‌ترین همسایگی (K-NN)[2] ، آنالیز مجزاساز خطی (LDA)[3] را می‌توان نام برد. در [5] ترکیبی از مدل مخفی مارکوف و MLP استفاده شده است. [9] ، [10] ، [11] ، [12]، [18] ، [20] و [21] مدل مخفی مارکوف، [16] و [27], [34] شبکه عصبی و در [35] از ماشین بردار پشتیبان[4] استفاده شده است.

۳-۳-۱ شبکه عصبی

روشی است که بر پایه اتصال به هم پیوسته چندین واحدپردازشی ساخته می‌شود. از تعدادی نرون تشکیل می‌شود که ورودی را به خروجی ربط می‌دهند. از روش‌هایی است که برای تشخیص دیداری صحبت به فراوانی به کار گرفته شده است در [27] از شبکه عصبی MLP با توابع انتقال تانژانت سیگموید در لایه پنهان و خروجی به استفاده شده است. این شبکه در دو مرحله، یک بار برای آموزش و تست فقط بر روی زاویه‌های گشودگی افقی و بار دیگر بر روی زاویه‌های گشودگی افقی و عمودی به کار گرفته شده است. در [34] نیز از شبکه عصبی MLP استفاده شده که در آن تابع فعال‌سازی ورودی و خروجی به صورت خطی و تابع فعال‌سازی لایه میانی سیگموید می‌باشد و خروجی به صورت

[1] Neural Networks
[2] K- Nearest Neighbor
[3] Linear Discriminate Analysis
[4] Support Vector Machine

$Y=W_2 *F(W_1 *X + B_1) + B_2$ می‌باشد و وزن‌های شبکه طوری تغییر می‌کنند که مجموع مربع خطا می‌نیمم شود. در [36] شبکه چند لایه Feed Forward Back Propagation Error به کار گرفته شده است.

۳-۳-۱-۱ شبکه‌های پیش‌خور

شبکه‌های پیش‌خور، شبکه‌هایی هستند که مسیر پاسخ آن‌ها همواره رو به جلو پردازش می‌شود و به نرون‌های لایه‌های قبل باز نمی‌گردد. در این نوع شبکه‌ها به سیگنال ها اجازه می‌دهند تنها از مسیر یک طرفه عبور کنند یعنی از ورودی تا خروجی. بنابراین باز خوردی وجود ندارد یعنی که خروجی هر لایه تاثیری بر همان لایه ندارد.

۳-۳-۱-۲ الگوریتم پس انتشار خطا

عمده‌ترین کاربرد قانون یادگیری پس انتشار، در شبکه‌های عصبی پیش‌خور است که عموما شبکه‌های چند لایه پرسپترون[1] هم نامیده می‌شوند. این الگوریتم بر قانون یادگیری اصلاح خطا مبتنی می‌باشد. این قانون از دو مسیر اصلی تشکیل شده است. مسیر اول یا مسیر رفت است که در این مسیر، بردار ورودی به شبکه اعمال و تاثیرش از طزیق لایه میانی به لایه خروجی انتشار می‌یابد. در این مسیر پارامترهای شبکه بدون تغییر در نظر گرفته می‌شوند. در مسیر دوم یا مسیر برگشت پارامترهای شبکه تغییر کرده و تنظیم می‌شوند. این تنظیم مطابق با قانون اصلاح خطا صورت می‌گیرد.

۳-۳-۲ مدل مخفی مارکوف

در [37] از سیستم HMM از ۳۳ مدل HMM برای شناسایی ۳۳ کلمه تشکیل شده است. هر مدل HMM یک مدل ۳ حالته چپ به راست با ۲ مخلوط گوسی[2] برای هر حالت می‌باشد. ابتدا مدل ها

[1] Perceptron
[2] Gaussian Mixture

مقداردهی اولیه شده و سپس با نسخه جاسازی شده[1] آموزشی از الگوریتم بام- ولش [2] دوباره تخمین زده می‌شوند. در ادامه داده آموزشی هم‌تراز شده برای مدل شدن با الگوریتم ویتربی[3] برای محاسبه چگالی زمانی حالات مورد استفاده قرار گرفته است. برای شناسایی یک کلمه جدید ، ویژگی‌های استخراج شده از آن، به عنوان ورودی به سیستم HMM اعمال و ماکزیمم احتمال مدل به عنوان خروجی شناسایی شده و کلمه متناظر به شکل متن نمایش داده شده است.

[1] Embedded
[2] Baum–Welch
[3] Viterbi

فصل چهارم : ویژگی‌های استخراجی و پیاده‌سازی روش پیشنهادی و معرفی پایگاه داده

۴-۱ پایگاه داده

در این کار از پایگاه داده‌ای که در [27] به کار برده شده است استفاده نمودیم. که در آن از چند مونث و مذکر برای ادای تعدادی از کلمات فارسی استفاده شده است که هر گوینده دو بار یا بیشتر کلمات را ادا کرده است. تصویر چهره از قسمت پایین صورت گوینده‌ها می‌باشد. از ۵ زن و ۱ مرد در این کار استفاده نمودیم که هر کدام کلمات تک سیلابی را ۲ و یا ۳ بار تکرار نمودند. تصاویر در اندازه ۳۲۰*۲۴۰ می‌باشند. این پایگاه داده شامل فایل‌های صوتی نیز می‌باشد که با توجه به عدم نیاز به آن‌ها مورد استفاده قرار نگرفتند. این مجموعه تمامی مصوت‌ها را در بر می‌گیرد در زیر کلمات تک سیلابی ادا شده آورده شده است.

جدول۴-۱ کلمات تک سیلابی در بانک اطلاعاتی

أ	إ	اُ	آ	ای	او
هشت	یک	دو	آب	سی	دور
ده	سه	نه	آش	دیر	دوغ
صد	شش	سر	چای	سیب	سور
سر	سر	در	سار	شیر	سوپ
اسب	دل	گرگ	دار	زیر	موش
قند	خرس	موز	گاو		رو
سد	کیک	تند	مار		

هرچند علاوه بر این‌ها خود مصوت‌ها نیز به تنهایی تلفظ شده و در این بانک موجود می‌باشد.

۴-۱-۱ جداسازی ویدیوهای ضبط شده

پایگاه داده مورد نظر به صورت ویدیویی می‌باشد که هر ویدیو دارای فریم‌های متفاوتی می‌باشد در این کار از تعداد مشخصی فریم‌های ویدیو برای استخراج ویژگی استفاده شده است. هر فریم تصویر دارای بعد ۳۲۰*۲۴۰ و رنگی می‌باشد. ویدیوهای مربوط به تلفظ هرکلمه را با استفاده از نرم افزار Ulead Video Studio جدا می‌کنیم که کلمات به صورت تک سیلابی می‌باشند. و با استفاده از نرم افزار Ulead GIF Animator می‌توانیم این ویدیوها را به رشته تصاویر تبدیل کنیم.

۴-۲ ویژگی‌های استخراج شده

با توجه به اینکه ویدیوهای موجود در پایگاه داده‌ای که استفاده نمودیم دارای فریم‌های متفاوتی بود در گام نخست پس از تبدیل ویدیوها به فریم‌های تصویرشان، با توجه به اینکه کلمات ادا شده دارای تعداد فریم‌های متفاوتی می‌باشند تعداد ۳۰ فریم از هر ویدیو را که مربوط به تلفظ مصوت‌های موجود در کلمه می‌باشد را انتخاب نمودیم. از میان کلمات ادا شده ۵۸۰ کلمه تک سیلابی را انتخاب و ضرایب مل فرکانسی که در بخش‌های بعدی مفصل شرح داده شده است را استخراج کردیم. برای محاسبه ضرایب ابتدا هر فریم تصویر ورودی را از ماتریس ۲ بعدی به یک بردار یک بعدی تبدیل و بعد از آن الگوریتم استخراج ویژگی را به بردار مورد نظر اعمال کردیم. خروجی از مرحله استخراج ویژگی ماتریسی است که شامل بردارهای ویژگی بدست آمده از تمام فریم ها می‌باشد.

بعد از فریم‌بندی و پنجره‌گذاری به طول ۲۵۶ با هم‌پوشانی ۱۰۰ نمونه، برای هر بلوک فریم تصویر ۱۳ ضریب MFCC استخراج شد. ردیف‌هایی از ماتریس متناظر با تعداد بلوک‌های فریم ویدیو، در حالیکه ستون‌ها متناظر با ضرایب بردار ویژگی است. این ویژگی‌ها رابرای ۵۸۰ ویدیو محاسبه کردیم. در نهایت خروجی را به عنوان ورودی به شبکه عصبی اعمال کردیم. از شبکه عصبی با ۲۰ نرون میانی

دارای تابع فعالسازی تانژانت سیگموید[1] در لایه میانی و خطی در خروجی استفاده کردیم که برای آموزش روش گرادیان نزولی با نرخ آموزش متغیر را به کار گرفتیم که نتایج خوبی حاصل نشد. مشاهده می‌کنیم که ابعاد ویژگی‌های استخراج شده بسیار بزرگ می‌باشد که سبب به وجود آمدن پیچیدگی و نیز کاهش سرعت محاسبات و کاهش سرعت شبکه می‌شود. بنابراین تصمیم به یافتن ناحیه‌ای شامل لب گرفتیم که هم بتوانیم سایز تصاویر را کاهش دهیم و هم از ورود اطلاعات اضافی که مربوط به نواحی صورت در اطراف دهان می‌باشد به بردار ویژگی‌هایمان جلوگیری کنیم. چون تصاویر از نیمه پایینی صورت است پس به دنبال روشی برای استخراج ناحیه‌ای که تنها دهان را شامل شود برآمدیم.

۴-۳ جداسازی ناحیه لب

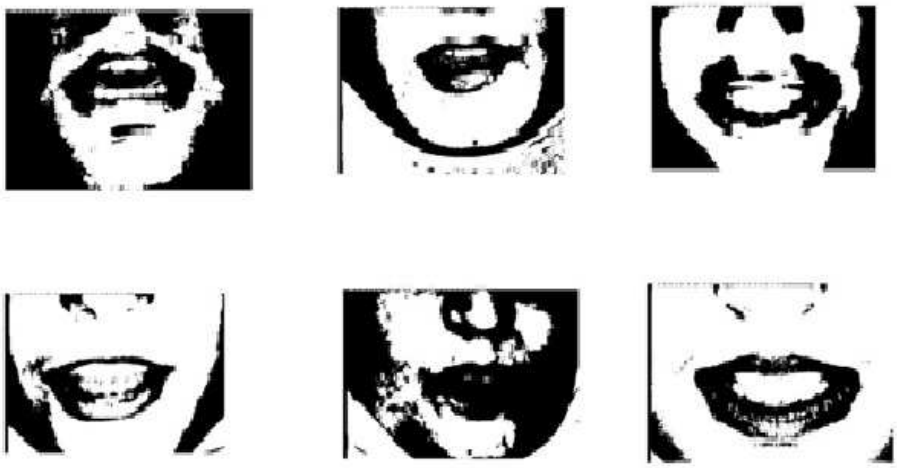
برای جداسازی ناحیه لب روش‌های مختلفی به کار گرفتیم که هر کدام از آن‌ها و نتایج حاصل از آن در ادامه بیان شده است.

۴-۳-۱ آستانه گذاری[2]

یکی از روش‌ها برای قسمت‌بندی[3] تصاویر ، آستانه‌گذاری می باشد اعمال این روش بر روی تصاویر به صورت زیر انجام شد.

[1] Tangent Sigmoid
[2] Thresholding
[3] Segmentation

شکل ۴-۱ آستانه گذاری با ترشلد ۰.۴



شکل ۴-۲ آستانه گذاری با ترشلد ۰.۵

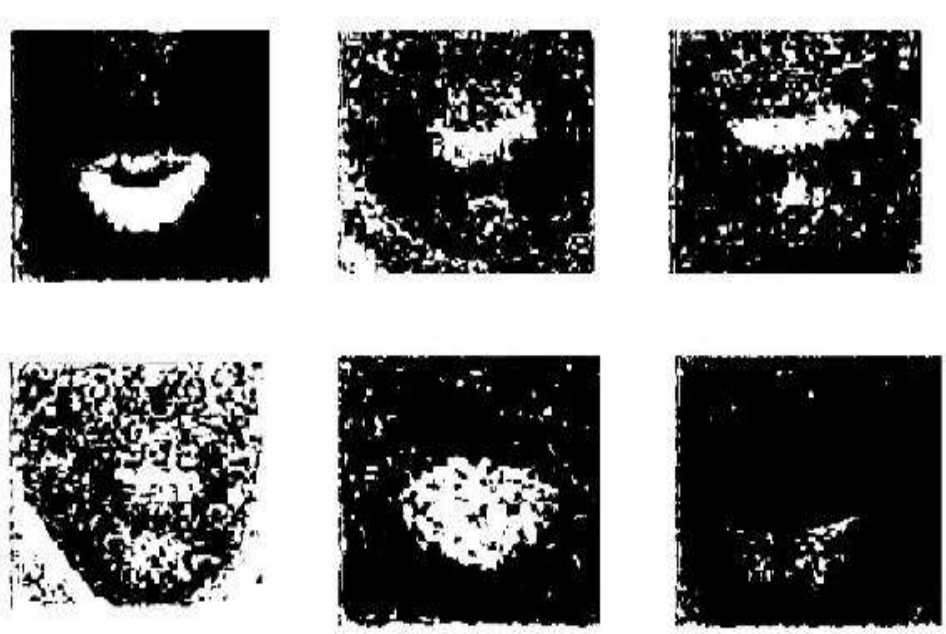
برای تصویر هر فرد در یک آستانه متفاوت ناحیه لب ظاهر می‌شود بنابراین استفاده از این روش با یک آستانه مشخص منجر به دستیابی به ناحیه لب برای تمام تصاویر نخواهد شد.

۴-۳-۲ استفاده از روش حذف رنگ قرمز

این روش روی مقادیر رنگ آبی و سبز تمرکز می‌کند و با استفاده از نسبت مؤلفه سبز به آبی ناحیه لب را استخراج می‌کند. ابتدا تصویر را با یک فیلتر گوسی کانوال کردیم تا نویز حذف شود. سپس نسبت رنگ‌ها را توسط مقیاس لگاریتمی به دست آوردیم.

$$log \frac{G(i,j)}{B(i,j)} \leq \beta$$ رابطه (۴-۱)

مقیاس لگاریتمی کنتراست را افزایش می‌دهد و با تغییر مقدار ترشلد β ناحیه دهان شناسایی می‌شود. اعمال الگوریتم فوق به صورت زیر انجام گرفت و با تغییر مقدار β نیز نتایج مطلوبی حاصل نگردید.

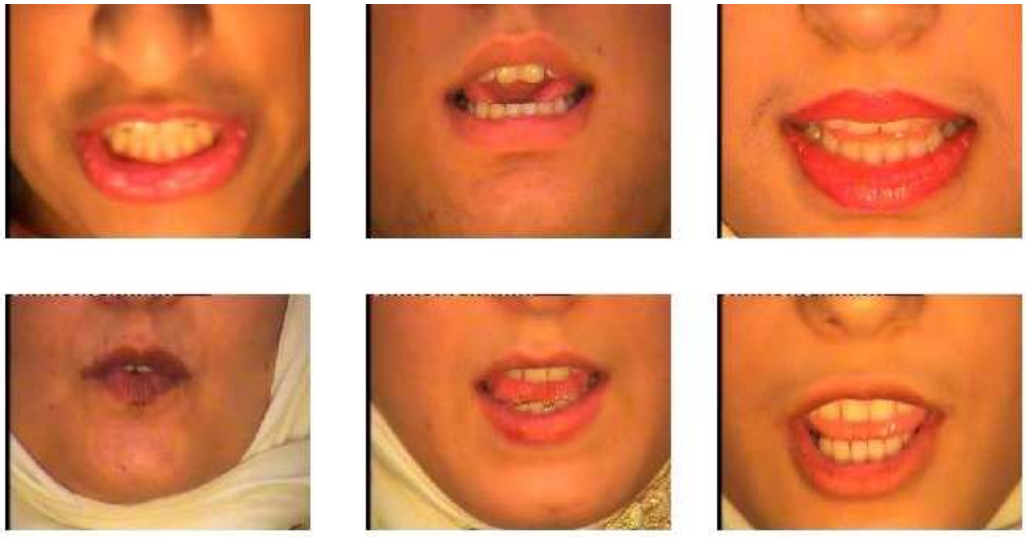


شکل ۴-۳ استفاده از الگوریتم حذف رنگ قرمز با β=۰.۵

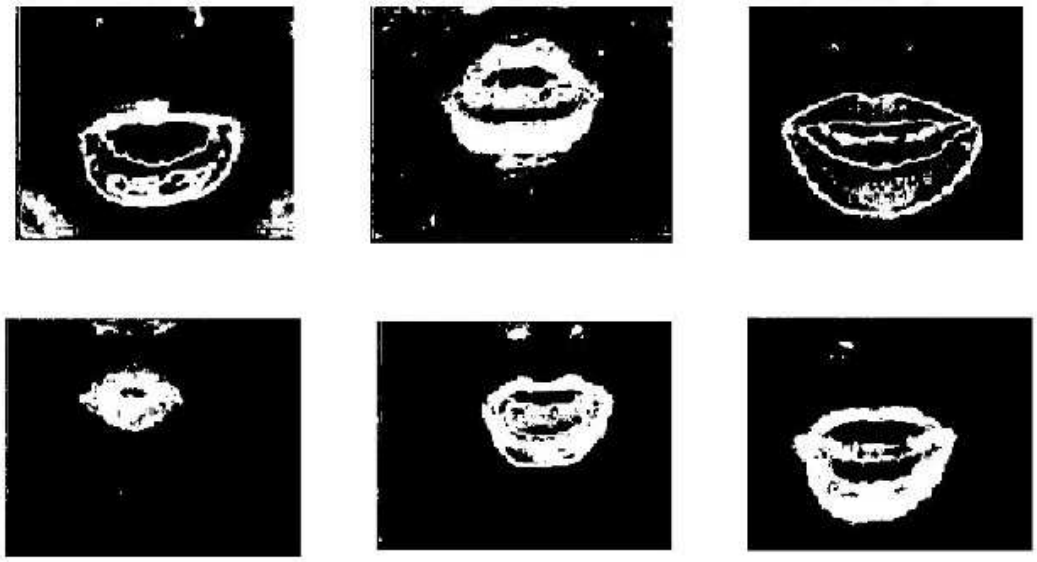
۴-۳-۳ آنالیز ترکیب رنگ لب وپوست

برای استفاده از این روش ابتدا تصاویر با یک فیلتر گوسی با پنجره‌ی ۳*۳ و سیگمای ۱۰ کانوال کردیم و سپس با روش سعی و خطا مقادیر آستانه بالا وپایینی برای تصاویر در نظر گرفتیم. که این مقادیر $(L_{lim}, U_{lim}) = (۲.۱و۳.۲۵)$ در نظر گرفتیم و با استفاده از این مقادیر و استفاده از روش برچسب‌گذاری و بکارگیری دستورات مورفولوژی همچون گشودگی ناحیه لب و با استفاده از نسبت مولفه های قرمز به سبز از تصویر که توسط رابطه (۴-۲) نشان داده شده است به شکل زیر استخراج شد.

$$h = \frac{R(i,j)}{G(i,j)}$$ رابطه (۴-۲)



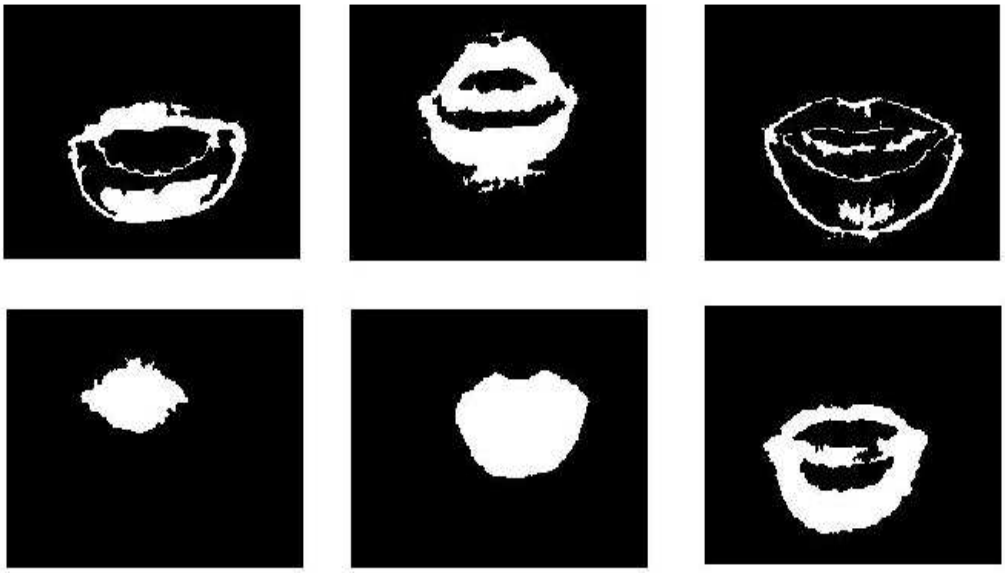
شکل ۴-۴ تصاویر مربوط به گوینده ها



شکل ۴- ۵ شکل لب استخراج شده بعد از اعمال الگوریتم

۴-۳-۴ برچسب‌گذاری اجزا

ناحیه دهان را از تصاویر منتج شده از قسمت قبل با برچسب‌گذاری و اعمال عملیات شکل شناسی مجزا می‌کنیم. در این کار از دستوراتی همچون Imopen , Imerode استفاده نمودیم. برای برچسب-گذاری از اولین سطر شروع کرده و در طول سطر حرکت می‌کنیم تا به اولین پیکسل با مقدار ۱ برسیم این پیکسل به عنوان عضوی از مجموعه ۱ علامت گذاری می‌شود. همه‌ی همسایگی‌های آن در سطرها و ستون‌ها بررسی می‌شود در صورتی که آن‌ها نیز دارای همین مقدار باشند عضوی از مجموعه محسوب می‌شوند در غیر این صورت اگر این پیکسل دارای مقدار ۱ باشد و با پیکسل قبلی که با مقدار ۱ علامت‌گذاری شد در همسایگی نباشد این پیکسل جدید و تمام پیکسل‌های همسایه آن که دارای مقدار ۱ هستند با ۲ علامت‌گذاری شده به این ترتیب تمام مؤلفه‌های درون تصویر بر چسب-گذاری می‌شوند که هر یک دارای مساحت مشخصی می‌باشند که بیشترین مساحت مربوط به لب می‌باشد که با مشخص نمودن مقدار ماکزیمم برای این اجزا می‌توان لب را استخراج نمود.



شکل ۴- ۶ شکل لب استخراج شده بعد از برچسب‌گذاری

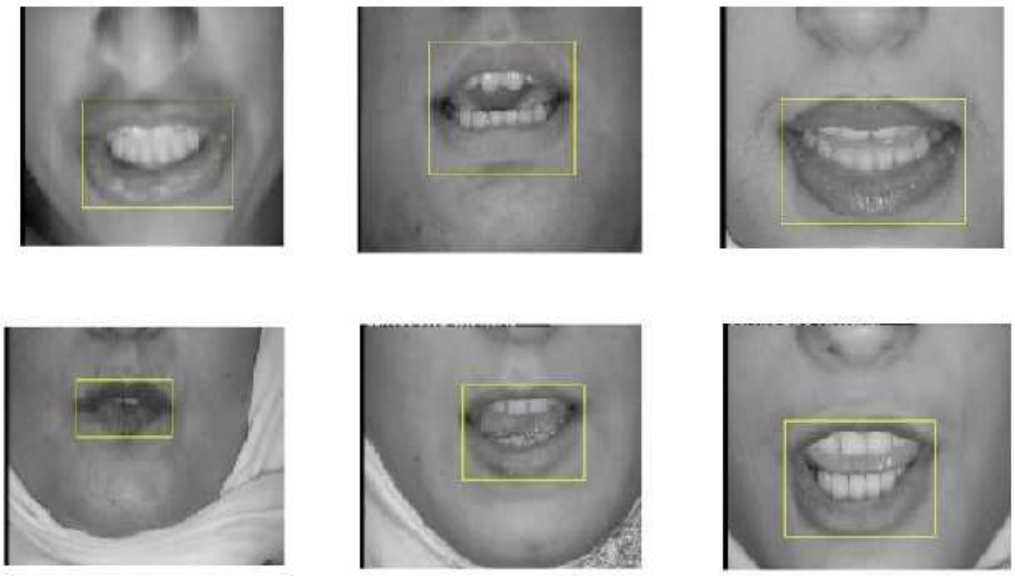
در این روش در تصاویر برخی از گویندگان ناحیه لب به خوبی استخراج شد اما در برخی دیگر بعضی از پیکسل‌های مجاور نیز به عنوان ناحیه لب آشکار شد. هرچند برای یافتن یک ناحیه مطلوب حول لب الگوریتم بالا کافی و مناسب است اما برای یافتن مقادیر دقیق پهنا و ارتفاع نیاز به استخراج دقیق لب خواهیم داشت.

۴-۳-۵ جعبه محاطی[1]

برای تصاویر شکل (۴-۶) جعبه‌ای که تصویر مورد نظر را احاطه می‌کند محاسبه می‌کنیم و بعد از اینکه این ناحیه مستطیلی به دست آمد ویژگی‌های آن را مانند ضرایب مل فرکانسی استخراج می-کنیم. برای این کار بعد از انجام مراحل قبل و یافتن ناحیه مربوط به دهان ابتدا برای هر ناحیه مرکز آن ناحیه را پیدا می‌کنیم. بعد مختصات افقی نقطه وسط از تصویر لب را با یافتن می‌نیمم مقدار مجموع سطرهای تصویر محاسبه و مختصات عمودی آن را نصف سایز عمودی تصویر قرار می‌دهیم و

[1] Bounding Box

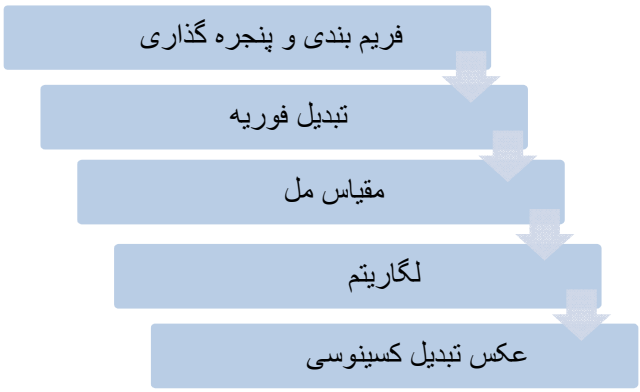
بعد با تعیین کمترین فاصله بین مرکز و این نقطه ، مختصات مستطیلی که این ناحیه را شامل می-شود به دست می‌آوریم . که از این مختصات برای یافتن پهنا و ارتفاع دهان می‌توانیم استفاده کنیم اما با توجه به اینکه با این روش در مورد برخی از تصاویر ناحیه درست استخراج نشده و یا شامل قسمت-های اضافی از صورت بود روش دقیقی برای محاسبه این ویژگی‌ها نبود.



شکل ۴-۷ مستطیل محاطی لب

۴-۴ ضرایب مل فرکانسی

از ضرایب مل فرکانسی به وفور برای شناسایی صوت استفاده شده است و این ضرایب از سیگنال صوتی استخراج شده و به عنوان بردار ویژگی بکار برده شده است [38] . از این ضرایب برای شناسایی تصویر و شناسایی چهره در [39] و [40] استفاده شده است. و همچنین برای شناسایی اثر انگشت و شناسایی حرکات دست از تصویر آنها در [34] , [35] و برای شناسایی تصاویر کف دست در [36] استفاده شده است. در شکل زیر الگوریتم محاسبه این ضرایب نشان داده شده است.



شکل ۴-۸ مراحل محاسبه ضرایب مل فرکانسی

۴-۴-۱ فریم‌بندی

چون MFCC روی سیگنا ل‌های یک بعدی کار می‌کند پس در اولین گام باید تصویر مورد نظر را که دو بعدی می‌باشد را به سیگنالی یک بعدی تبدیل کنیم. این کار را می‌توان توسط ذخیره تمام سطرها پشت سر هم و یا ذخیره تمام ستون‌ها پشت سر هم انجام داده و سیگنالی یک بعدی شکل داد. در گام دوم باید فریم‌بندی روی سیگنال صورت داد و سیگنال را به گروه‌هایی با تعداد مشخصی از نمونه-ها دسته‌بندی کرد که بدین منظور سیگنال یک بعدی را به فریم‌های کوچکی که شامل ۲۵۶ نمونه می‌باشد تقسیم می‌کنیم برای حفظ اطلاعات سیگنال ، هر فریم تقسیم شده با فریم قبلی آن باید هم-پوشانی داشته باشد به همین منظور تعداد ۱۰۰ نمونه از هر فریم با فریم قبلی هم‌پوشانی خواهد داشت.

۲-۴-۴ پنجره‌گذاری

پنجره‌گذاری برای می‌نیمم کردن شکستگی‌ها و ناپیوستگی‌ها در ابتدا و انتهای هر فریم بکار برده می‌شود. که معمولاً برای این کار از پنجره همینگ[1] که رابطه آن در زیر بیان شده استفاده می‌شود که ما نیز همین پنجره را بکار گرفتیم. (N سایز پنجره می‌باشد).

$$W_H(n)=0.54-0.46\cos\left(\frac{2n\pi}{N-1}\right) \qquad n=0,1,\ldots,N-1$$ رابطه (۴-۳)

۳-۴-۴ تبدیل فوریه گسسته[2]

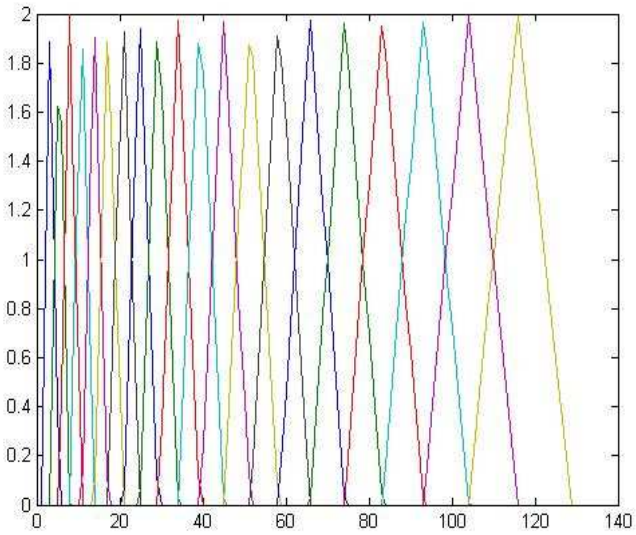
تبدیل فوریه برای تبدیل سیگنال از حوزه فضایی به حوزه فرکانسی استفاده و بنابراین هر فریم به حوزه فرکانسی تبدیل شده و اندازه اسپکتروم[3] محاسبه می‌شود.

$$S[k]=\sum_{n=0}^{N-1} s[n]e^{-j2\pi\frac{k}{N}n}$$ رابطه (۴-۴)

۴-۴-۴ مقیاس مل [4]

مل فرکانسی وارپینگ توسط یک بانک فیلتر مل که مجموعه‌ای از فیلترهای میان‌گذر با پهنای باند ثابت و فاصله‌گذاری روی مقیاس مل می‌باشد انجام می‌شود. هر فیلتر پاسخ فرکانسی میان گذر مثلثی دارد که این فیلترها روی تمام محدوده فرکانسی از صفر تا فرکانس نایکوئیست[5] جدا شده‌اند. تعداد فیلترها یکی از پارامترهایی است که روی دقت شناسایی تاثیر دارد.

[1] Hamming Window
[2] Discrete Fourier transform
[3] Spectrum
[4] Mel Scale
[5] Nyquist



شکل ۴-۹ فیلتر بانک مثلثی

با استفاده از رابطه زیر فرکانس بر حسب هرتز را می‌توان به فرکانس مل تبدیل کرد.

$$mel(f)= 2595*\log10(1+f/700)$$ رابطه (۴-۵)

با کمک فیلتر بانک انرژی در هر نقطه محاسبه و لگاریتم آنها مل کپستروم‌ها را که برای محاسبه ضرایب مل ضروری است را ایجاد می‌نماید.

در [34] از MFCC برای شناسایی اثر انگشت استفاده شده که این ضرایب از تصاویر استخراج شده و به همراه ضرایبی که پس از تبدیل موجک گرفتن از تصویر به دست آمده ، ضرایب چند جمله‌ای آنها را محاسبه و به عنوان ورودی شبکه عصبی در نظر گرفته شده است. چون MFCC ها به عدم تطابق کانال بین آموزش و تست حساسند ضرایب چند جمله‌ای به آن‌ها اضافه می‌شوند. اهمیت این ضرایب به این دلیل است که آن‌ها می‌توانند اطلاعات مهم را همچون متوسط[1] وشیب[2] و مقدار انحناء[3] در

[1]Mean
[2] Slope
[3] Curvature

مورد شکل یک تابع زمانی را حفظ کنند. در این تحقیق برای استفاده از ضرایب چند جمله‌ای، توابع زمانی از ضرایب کپسترال توسط نمایش چند جمله‌ای متعامد در ۹۰ میلی ثانیه با گام ۱۰ میلی ثانیه بسط داده شده است. که این مدت زمان ۹۰ میلی ثانیه‌ای به نظر می‌رسد که برای حفظ اطلاعات انتقالی مناسب باشد. در نهایت ضرایب کپسترال به همراه ضرایب چند جمله‌ای مرتبه اول و ضرایب چند جمله‌ای مرتبه دوم استفاده شده است. مرحله کلاسه‌بندی در سیستم‌های شناسایی اتوماتیک در واقع یک فرایند تطبیق ویژگی بین ویژگی‌ها از تصاویر جدید اثر انگشت می‌باشد. برای انجام این کار از شبکه عصبی MLP استفاده شده و الگوریتم پس انتشار خطا برای آموزش به کار گرفته شده است. در [35] از تصاویری با زمینه ساکن[1] که یا تیره و یا روشن می‌باشد و حالت‌های مختلف دست را نشان می‌دهد استفاده شده است. ۱۳ ضریب MFCC از تصاویر سطح خاکستری محاسبه شده و به عنوان ورودی به کلاسه‌بند SVM اعمال شده است. در این کار کلاسه‌بندی بین ۱۰ کلاس مختلف انجام شده و نرخ شناسایی هر کلاس محاسبه شده است. در [36] برای شناسایی کف دست[2] ضرایب MFCC به کار گرفته شده و بین ۱۲ تا۲۰ ضریب استخراج و به ضرایبی که بعد از تبدیل ویولت گرفتن از تصاویر استخراج شده اضافه و به عنوان ورودی به شبکه عصبی مانند [34] اعمال شده است.

۴-۴-۵ تبدیل کسینوسی گسسته

در واقع این عمل برای برگرداندن به حوزه فضایی صورت می‌گیرد و ضرایب مل فرکانسی در نتیجه آن حاصل می‌شود.

$$C(n)= \sqrt{\frac{2}{N}} \sum_{k=1}^{N} \log(s_k) \cos\left(\frac{n\pi}{N}(k-0.5)\right)$$ رابطه (۴-۶)

که S_k خروجی از k امین فیلتر و N تعداد فیلترها و n= 0 , 1 , …. , M-1 و M تعداد ضرایب است. که همان طور که مشاهده می‌شود بعد از گرفتن لگاریتم از خروجی فیلتر بانک، عکس تبدیل

[1] Static
[2] Palmprint

کسینوسی اعمال و در نهایت ضرایب مل فرکانسی منتج می‌شود. گرچه DFT معمولا برای آنالیز کپستروم استفاده می‌شود اما از آن جایی که DCT برای فشرده سازی به کار برده می‌شود اطلاعات بیشتری را در تعداد کمتری از ضرایب متمرکز می‌کند. بنابراین فضای کمتری را برای نمایش ضرایب کپستروم استفاده می‌کند و چون مقادیر مهم انرژی را شامل می‌شود نسبت به DFT مناسب‌تر است.

۴-۴-۵-۱ محاسبه ضرایب کسینوسی و ویولت

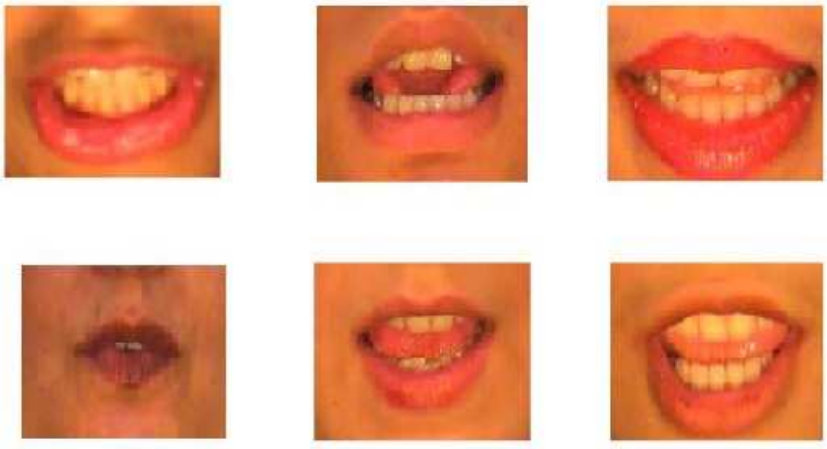
از جمله ویژگی‌هایی که استخراج نمودیم ضرایب کسینوسی و ویولت بود. که ماتریس ضرایب کسینوسی پس از استخراج توسط اسکن زیگزاگ به بردار تبدیل کردیم و ضرایب مختلفی از این ضرایب را انتخاب نمودیم و همچنین از ۱/۲ و ۱/۴ و ۱/۸ بردار ضرایب کسینوسی برای محاسبه ضرایب MFCC استفاده کردیم . ضرایب اصلی ویولت و نیز ضرایب MFCC استخراج شده از این ضرایب را نیز به عنوان ویژگی در نظر گرفتیم.

۴-۴-۵-۲ محاسبه ضرایب مل فرکانسی

ناحیه مستطیلی را از تمام فریم‌های ویدیو استخراج می‌کنیم با توجه به اینکه پهنا و اندازه لب گویندگان در فریم‌های مختلف هنگام تلفظ کلمات تغییر می‌کند ابعاد مستطیل متفاوت بوده و بعد از استخراج ضرایب مل فرکانسی هر تصویر تعداد این ضرایب یکسان نمی‌باشد بنابراین مجبوریم که می-نیممی برای تعداد این ضرایب در نظر گرفته و فقط این مقدار از ضرایب را به عنوان بردار ویژگی در نظر می‌گیریم اما چون تغییرات سایز این جعبه مستطیلی شکل باعث می‌شود که قسمتی از اطلاعات را نادیده بگیریم تصمیم به یافتن ناحیه‌ای مطلوب با اندازه‌ای مشخص برای تمام گویندگان گرفتیم. ما ضرایب مل فرکانسی را برای این ناحیه مستطیلی محاسبه نمودیم که نتایج خوبی حاصل نگشت بنابراین به دلایل فوق، ناحیه‌ای حول لب‌ها استخراج کردیم.

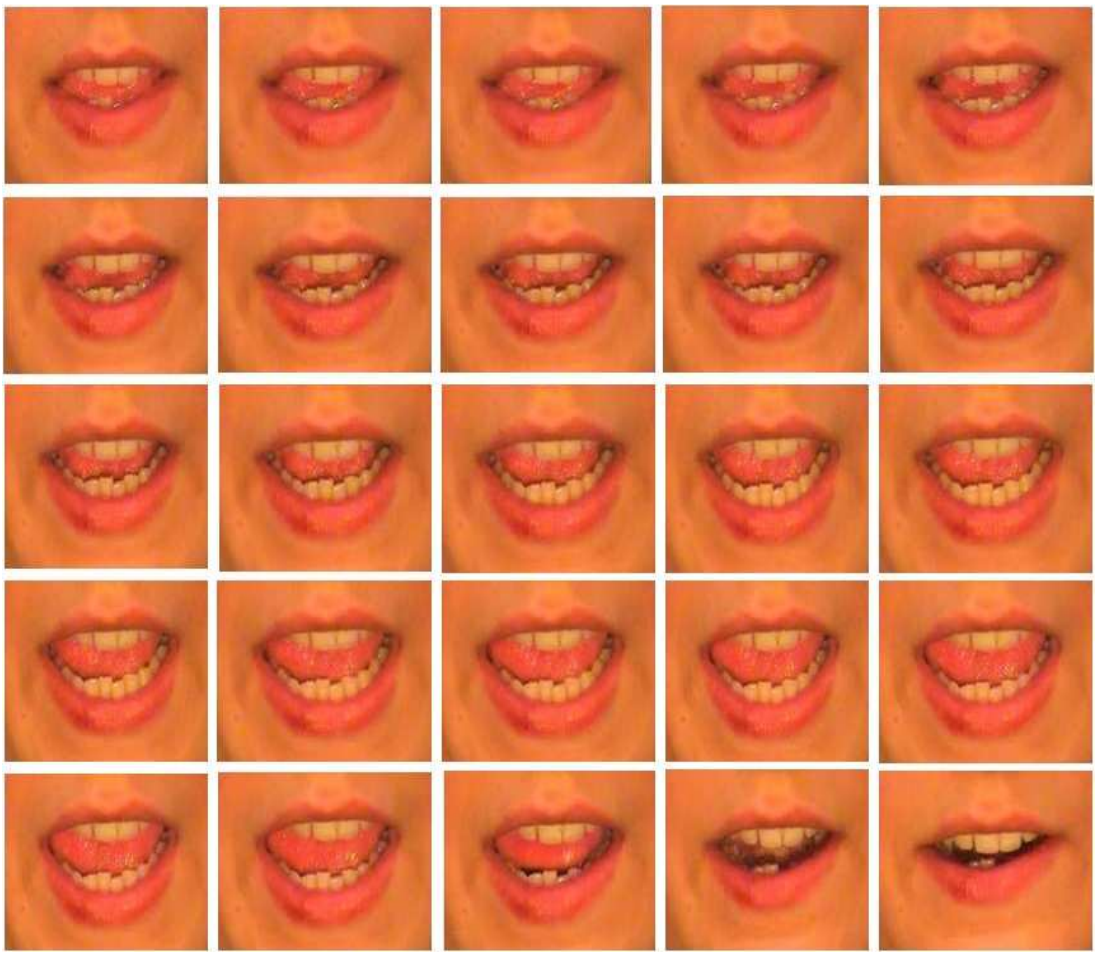
۴-۵ یافتن مرکز لب و استخراج ناحیه ای حول لب

با استفاده از ناحیه به دست آمده با الگوریتم بالا با مشخص نمودن مرکز آن ناحیه‌ای را در اطراف لب استخراج کردیم تا بتوانیم سایر قسمت‌های اضافی از تصاویر را از آن‌ها جدا نماییم و با کاهش سایز تصویر عملکرد آن‌ها را بهبود بخشیم. در نتیجه سایز تصویر را به ۲۲۰*۱۵۰ تغییر دادیم که این عمل را با استفاده از مراحل قبل و یافتن مرکز ناحیه استخراج شده از تصویر و مشخص نمودن طول و عرض مشخص، انجام دادیم. با توجه به اینکه پهنای لب و شکل لب هر گوینده و فاصله صورت تا دوربین متفاوت است و اینکه ما می‌خواهیم سایز تصویر نهایی برای تمام افراد یکسان باشد این ناحیه برای برخی گویندگان با لب کوچک ، نواحی اضافه‌تری در اطراف لب نسبت به گویندگانی که لب‌های بزرگتری دارند شامل می‌شود اما این اندازه حداقل اندازه‌ای است که به ازای آن تمام لب را شامل می‌شود. با استفاده از نتایج به دست آمده در بخش (۴-۳-۴) ناحیه مورد نظر[1] به صورت زیر استخراج می‌شود.



شکل ۴-۱۰ ناحیه مورد نظر پیرامون لب

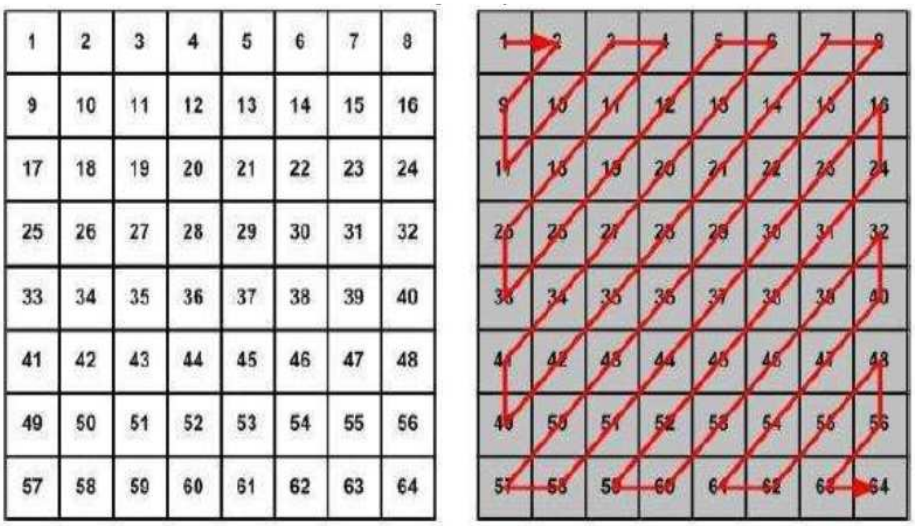
[1] Region Of Interest



شکل ۴-۱۱ تعداد ۲۵ فریم مربوط به کلمه خرس بعد از یافتن ناحیه مورد نظر

۴-۵-۱ اسکن زیگزاگ

اسکن زیگزاگ مطابق با شکل (۴-۱۲) صورت می‌گیرد که باعث دستیابی به ضرایب مهم می‌شود. ضرایب مهم DCT عموماً در گوشه چپ و بالای ماتریس DCT یافت می‌شود با اسکن زیگزاگ احتمال دستیابی به این ضرایب افزایش می‌یابد چون ضرایب را به صورت نزولی مرتب می‌کند. با این روش ماتریس ضرایب DCT به برداری به ابعاد یک در حاصلضرب تعداد سطرها در ستون‌ها تبدیل و ضرایب فرکانس پایین در بالای بردار جمع می‌شوند. با این عمل مولفه‌های فرکانس بالا که در اثر نویز به وجود می‌آیند حذف می‌شوند و بردار ویژگی‌ها ناهمبسته می‌شود.



شکل ۴-۱۲ نحوه اسکن زیگزاگ ماتریس

با توجه به این که بزرگ بودن ابعاد ویژگی‌ها سبب می‌شود که اطلاعات اضافی زیاد شده و در نتیجه روند تصمیم‌گیری به خوبی صورت نگیرد. بنابراین برای کاهش ابعاد ویژگی‌ها باید از روش‌های کاهش ویژگی استفاده کنیم و سایز بردار ویژگی‌ها را کم کنیم.

برای این کار روش‌های مختلفی همچون PCA و LDA و LSDA [1] وجود دارد که از روش آخر استفاده کردیم و بردار ویژگی کاهش یافته با این روش را، به عنوان ورودی به شبکه اعمال کردیم و درصد شناسایی شبکه را محاسبه نمودیم.

۴-۵-۲ کاهش ویژگی با LSDA

چون تغییر حرکات لب به صورت نرم است در نظر گرفتن جداسازی بین کلاس‌های مختلف تنها کافی نمی‌باشد و اطلاعات ساختار مکانی نیز مهم می‌باشد. بنابراین چون LSDA هر دو ساختار جداسازی و هندسی داده‌ها را با هم در نظر می‌گیرد روش بسیار خوبی برای کاهش ویژگی می‌باشد. در [41] این روش معرفی شده است. ما نیز برای کاهش ابعاد ویژگی‌ها این روش را به کار می‌گیریم تا ببینیم با اعمال این روش دقت شناسایی سیستم چه تغییری می‌کند. با استفاده از این روش و

[1] Locality Sensitive Discriminant Analysis

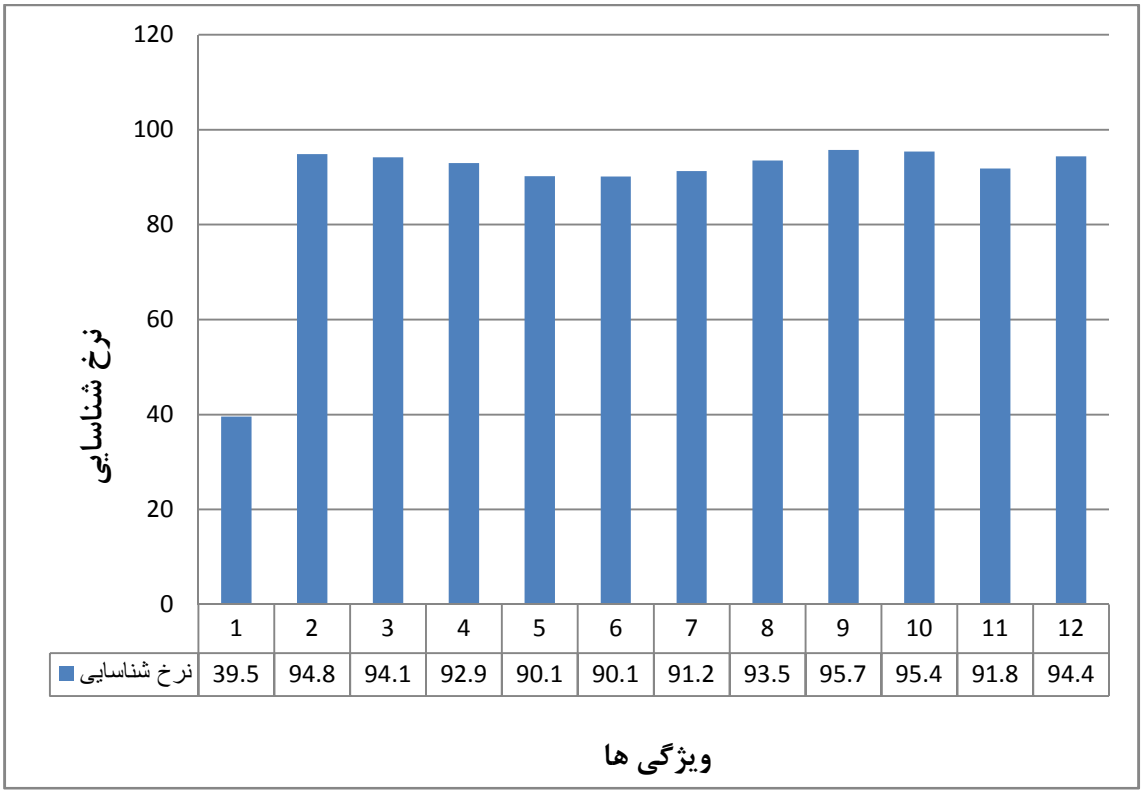
محاسبه بردارهای ویژه، اندازه ویژگی‌ها را به ۲۵ تغییر می‌دهیم و این ویژگی‌های جدید را به شبکه اعمال می‌کنیم.

۲۵ فریم از تصاویر را به صورت دستی انتخاب و تصاویر را با مقیاس ۰.۷ کوچک نمودیم برای این که اطلاعات کمتری از تصویر حذف شود و ابعاد ویژگی‌ها به گونه‌ای تغییر کند که بتوانیم آن‌ها را با کمک LSDA کاهش دهیم. تمام ویژگی‌هایی که قبلا بیان کردیم را با این روش کاهش سایز می-دهیم. در جدول زیر نتایج اعمال این روش کاهش ویژگی پس از ۵ بار آموزش و تست بیان شده است. از ۳۸۱ ویدیو برای آموزش و ۱۷۹ ویدیو را برای تست و ۲۰ ویدیو را برای اعتبار سنجی استفاده نمودیم. همان‌طور که قبلا هم بیان کردیم از شبکه عصبی Feed-Forward دو لایه با تابع فعالسازی تانژانت سیگموید در لایه اول و تابع خطی در لایه دوم استفاده نمودیم. ۲۰ نرون میانی و تعداد ۱۰۰۰ ایپاک[1] انتخاب و از گرادیان نزولی با نرخ آموزش متغیر برای آموزش شبکه استفاده نمودیم. ویژگی‌ها را به صورت زیر برچسب گذاری نمودیم.

۱-۱۰ضریب DCT پس از اسکن زیگزاگ، ۲-۵۰ ضریب DCT پس از اسکن زیگزاگ،

۳-۱۰۰ ضریب DCT پس از اسکن زیگزاگ، ۴- ۵۰۰ ضریب DCT پس از اسکن زیگزاگ، ۵- ۱۰۰۰ ضریب DCT پس از اسکن زیگزاگ، ۶ - کل ضرایب DWT، ۷- ضرایب MFCC از ماتریس DCT، ۸- ضرایب MFCC از ۱/۲ ضرایب DCT پس از اسکن زیگزاگ، ۹- ضرایب MFCC از ۱/۴ ضرایب DCT پس از اسکن زیگزاگ، ۱۰- ضرایب MFCC از ۱/۸ ضرایب DCT پس از اسکن زیگزاگ، ۱۱- ضرایب MFCC از ماتریس DWT، ۱۲- ضرایب MFCC از تصاویر.

[1] Epoch



شکل ۴-۱۳ نتایج حاصل از ویژگی‌ها + LSDA

۴-۵-۲-۱ استفاده از تابع Logsigmoid و تغییر الگوریتم آموزش

اگر به جای تابع خطی در خروجی از تابع Logsigmoid استفاده کنیم یا از این تابع هم در لایه میانی و هم در خروجی استفاده نماییم نتایج خوبی حاصل نمی‌شود. اگر الگوریتم آموزش را تغییر داده و از گرادیان نزولی همراه با ممنتوم[1] استفاده کنیم نیز نتایج مطلوبی به دست نمی‌آید.

۴-۵-۲-۲ استفاده از تابع Tansigmoid و الگوریتم ممنتوم

استفاده از تابع Tansigmoid در هر دو لایه و الگوریتم گرادیان نزولی با ممنتوم هم عملکرد خوبی نداشت.

[1] Momentum

مشاهده می‌کنیم که با استفاده از این روش نتایج خوبی حاصل می‌شود. این روش برای شناسایی دیداری صحبت استفاده شده است و نتایج حاصل از آن نسبت به سایر روش‌ها بهتر بوده است.

در [23] از DCT + LSDA برای استخراج ویژگی و کاهش ابعاد استفاده و با ویژگی‌هایی همچون DCT +LDA, DCT +PCA مقایسه شده است . در این کار نیز بعد از تبدیل کسینوسی با یک اسکن زیگزاگ برداری از سایز حاصلضرب تعداد سطرها در تعداد ستون‌ها خواهیم داشت چون ضرایب مهم در ابتدای این بردار هستند. که این ویژگی‌ها یک بار قبل از تنظیم نقاط انتهایی و بار دیگر بعد از آن به دست آورده شده‌اند. منظور از نقاط انتهایی نقاط مربوط به جداسازی سیلاب‌ها می‌باشد. بعد از جدا کردن سیلاب‌های کلمات توسط اطلاعات سیگنال صوتی مشخص شده است که برخی از نمونه‌ها اشتباه قطعه‌بندی شده‌اند بنابراین آن‌ها را به صورت دستی تنظیم می‌کنند. چندین ضریب اول آن‌ها در هر دو حالت انتخاب و به HMM اعمال شده است در زیر نتایج این روش آورده شده است.

جدول ۴- ۲ نتایج قبل از تنظیم نقاط انتهایی

تعداد ضرایب	۵۰	۶۰	۷۰	۸۰	۹۰	۱۰۰
DCT+PCA	٪۶۷.۷۱	٪۶۷.۷۱	٪۶۵.۶۳	٪۶۴.۵۸	٪۶۳.۵۴	٪۶۲.۵۰
DCT+LDA	٪۷۵	٪۷۵	٪۷۳.۹۶	٪۷۱.۸۸	٪۷۳.۹۶	٪۷۰.۸۳
DCT+LSDA	٪۷۲.۰۸	٪۷۹.۱۷	٪۸۰.۲۱	٪۸۰.۲۱	٪۷۸.۱۳	٪۷۶.۰۴

جدول ۴- ۳ نتایج بعد از تنظیم نقاط انتهایی

تعداد ضرایب	۵۰	۶۰	۷۰	۸۰	۹۰	۱۰۰
DCT+PCA	٪۶۷.۷۱	٪۶۷.۷۱	٪۶۵.۶۳	٪۶۴.۵۸	٪۶۴.۵۸	٪۶۳.۵۴
DCT+LDA	٪۸۰.۲۱	٪۸۱.۲۵	٪۸۱.۲۵	٪۷۹.۱۷	٪۸۲.۲۹	٪۸۱.۲۵
DCT+LSDA	٪۸۲.۲۹	٪۸۵.۴۲	٪۸۴.۳۸	٪۸۴.۳۸	٪۸۴.۳۸	٪۸۳.۳۳

در [27] که ما از پایگاه داده به کار رفته شده در این تحقیق استفاده نمودیم به کمک الگوریتم بیان شده در بخش (۳-۲-۶) شکل لب‌ها استخراج شده و سپس ویژگی‌هایی همچون پهنا و ارتفاع و زاویه گشودگی افقی و عمودی محاسبه شده و در نهایت با اعمال این ویژگی‌ها بیشترین نرخ شناسایی ۸۴٪ به دست آمده است.

ما در این تحقیق کارهای دیگری نیز انجام دادیم که در ادامه به آن‌ها اشاره می‌کنیم.

۴-۶ استخراج ویژگی از تصاویر مختلف

۴-۶-۱ استخراج ویژگی از تصاویر جدید

در تصاویر شکل (۴-۱۰) دیده می‌شود که در برخی از گویندگان ناحیه بزرگتری در نظر گرفته شده که این به اندازه لب‌ها و فاصله تا دوربین بر می‌گردد و در ضمن باید به کلمه تلفظ شده که باعث جمع شدن یا باز شدن دهان می‌شود نیز توجه نمود. پس تا این مرحله ناحیه مطلوب شامل دهان را استخراج نموده‌ایم. بعد از اینکه سایز تصاویر از ۳۲۰*۲۴۰ به ۲۲۰*۱۵۰ تغییر یافت با تبدیل تصاویر رنگی به تصاویر سطح خاکستری، از این تصاویر ضرایب مل فرکانسی (MFCC) و ضرایب کسینوسی (DCT) را استخراج می‌کنیم. همچنین تصاویر جدید را با فرمول زیر نرمالیزه نموده و این ویژگی‌ها را نیز برای آن محاسبه می‌کنیم. که σ انحراف استاندارد تصویر و μ میانگین تصویر است.

$$x_n = \frac{x-\mu}{6\times\sigma} + 0.5$$ رابطه (۴-۷)

۴-۶-۲ ضرایب مل فرکانسی و ضرایب کسینوسی

در این مرحله ضرایب مل فرکانسی و کسینوسی را از تصویر خاکستری اصلی و نیز از تصاویر نرمالیزه شده استخراج نمودیم با توجه به اینکه هر ویدیو مربوط به تلفظ کلمات تک سیلابی است و تعداد فریم‌های آن‌ها زیاد می‌باشد بنابراین پس از تبدیل ویدیو به رشته تصاویر ، به صورت دستی ۳۰ فریم

که در واقع شامل مصوت ادا شده می‌باشد را انتخاب کرده و پس آن را مجددا به ویدیو تبدیل نموده و از این ویدیوها که شامل ۳۰ فریم هستند استفاده می‌کنیم. با توجه به تعداد فریم‌ها و سایز تصاویر باز هم ابعاد ویژگی‌ها زیاد شد و در نتیجه بهبودی در نتایج حاصل نشد.

۴-۷ کاهش تعداد فریم‌ها و کاهش سایز تصاویر

با توجه به اینکه با استفاده از روش‌های کاهش ویژگی همچون LDA , PCA به دلیل خطایی که به دلیل بزرگ بودن بعد ویژگی‌ها به وجود می‌آمد امکان پذیر نشد بنابراین تعداد فریم‌ها را به ۲۰ فریم و سایز تصاویر را به ۱۸۰*۲۴۰ تغییر دادیم و ویژگی‌ها را با سایز و تعداد فریم کمتر محاسبه نمودیم.

۴-۷-۱ محاسبه ضرایب MFCC

بعد از یافتن ناحیه مطلوب از تصاویر جدید سایز به ۱۶۵*۱۳۰ تغییر یافت و ویژگی‌ها را از تصاویر سطح خاکستری و از نرمالیزه شده این تصاویر و همچنین از تصاویری که سایز آنها را به ۸۰ *۶۰ تغییر دادیم استخراج نمودیم. برای محاسبه ضرایب، فریم‌ها را به قسمت‌هایی از سایز ۲۵۶ با ۱۰۰ پیکسل هم‌پوشانی تقسیم نمودیم و ۱۳ ضریب MFCC از آن‌ها به دست آوردیم. در نهایت از هر فریم ۱۳*۲۱۲ ضریب و از فریم‌های با سایز ۸۰ *۶۰ تعداد ۱۳*۴۶ ضریب و در نهایت برداری از سایز ۵۵۱۲۰ و ۱۱۹۶۰ برای اعمال به شبکه حاصل شد.

۴-۷-۲ ضرایب DCT, DWT

ضرایب کسینوسی و ضرایب ویولت تصاویر را محاسبه و از ربع بالایی ماتریس ضرایب کسینوسی که شامل ضرایب بالا و چپ ماتریس که ضرایب مهم در این ماتریس هستند و از ماتریس ضرایب اصلی

ویولت ضرایب MFCC را محاسبه نمودیم. با استفاده از اسکن زیگزاگ[1] ماتریس ضرایب کسینوسی را به برداری از سایز حاصلضرب تعداد ستون‌ها و تعداد سطرها تبدیل کردیم و ۱۰۰۰ ضریب اول از این بردار را به عنوان بردار ویژگی در نظر گرفتیم.

ضرایب MFCC از ربع بالایی و سمت چپ ماتریس DCT که شامل ضرایب مهم از ضرایب کسینوسی است به دست می‌آوریم.

جدول ۴- ۴ نتایج حاصل از ویژگی های استخراجی از تصاویر اصلی با ۲۰ فریم

تصاویر اصلی	دقت شناسایی
۱۰۰۰ ضریب DCT	۲۶.۸۱
ضرایب DWT	۳۲.۳۹
ضرایب MFCC از DCT	۲۹.۹۷
ضرایب MFCC از DWT	۲۹.۴۱
ضرایب MFCC از تصاویر	۲۵.۱۳

جدول ۴- ۵ نتایج حاصل از ویژگی‌های استخراجی از تصاویر نرمالیزه شده با رابطه (۴-۷) با ۲۰ فریم

تصاویر نرمالیزه شده	نرخ شناسایی
۱۰۰۰ ضریب DCT	۳۰.۱۶
ضرایب DWT	۲۸.۱۱
ضرایب MFCC از DCT	۳۱.۶۵
ضرایب MFCC از DWT	۳۱.۲۸
ضرایب MFCC از تصاویر	۲۵.۸۸

[1] Zig-zag Scan

جدول ۴- ۶ نتایج حاصل از ویژگی های استخراجی از تصاویر کوچک شده با ۲۰ فریم

تصاویر ری سایز شده	نرخ شناسایی
۱۰۰۰ ضریب DCT	۲۵.۳۲
ضرایب DWT	۳۶.۸۶
ضرایب MFCC از DCT	۲۹.۴۱
ضرایب MFCC از DWT	۳۰.۷۲
ضرایب MFCC از تصاویر	۲۹.۰۶

در تمام جداول متوسط نرخ شناسایی پس از ۵ بار آموزش و تست محاسبه شده است.

یک بار دیگر این روند را ادامه دادیم ولی این بار بعد از اسکن فقط ۱۰ ضریب اول از ضرایب DCT را پس از اسکن زیگزاگ در نظر گرفتیم. در این حالت اندکی بهبودی حاصل شد اما مؤثر واقع نشد. نتایج در جداول زیر بیان شده است.

جدول ۴- ۷ نتایج حاصل از ۱۰ ضریب اول از ضرایب DCT تصاویر اصلی با ۲۰ فریم

تصاویر اصلی	نرخ شناسایی
۱۰ ضریب DCT	۳۵.۷۵
ضرایب DWT	۲۳.۳۴
ضرایب MFCC از DCT	۳۰.۶۹
ضرایب MFCC از DWT	۲۶.۰۵
ضرایب MFCC از تصاویر	۲۴.۹۰

جدول ۴- ۸ نتایج حاصل از ۱۰ ضریب اول از ضرایب DCT تصاویر نرمالیزه شده با ۲۰ فریم

تصاویر نرمالیزه شده	نرخ شناسایی
۱۰ ضریب DCT	۲۹.۴۹
ضرایب DWT	۱۷.۸۷
ضرایب MFCC از DCT	۲۷.۳۷
ضرایب MFCC از DWT	۲۶.۸۱
ضرایب MFCC از تصاویر	۲۲.۹۹

جدول ۴- ۹ نتایج حاصل از ۱۰ ضریب اول از ضرایب DCT تصاویر کوچک شده با ۲۰ فریم

تصاویر ری سایز شده	نرخ شناسایی
۱۰ ضریب DCT	۲۶.۱۲
ضرایب DWT	۲۰.۶۷
ضرایب MFCC از تصاویر	۲۶.۸۱

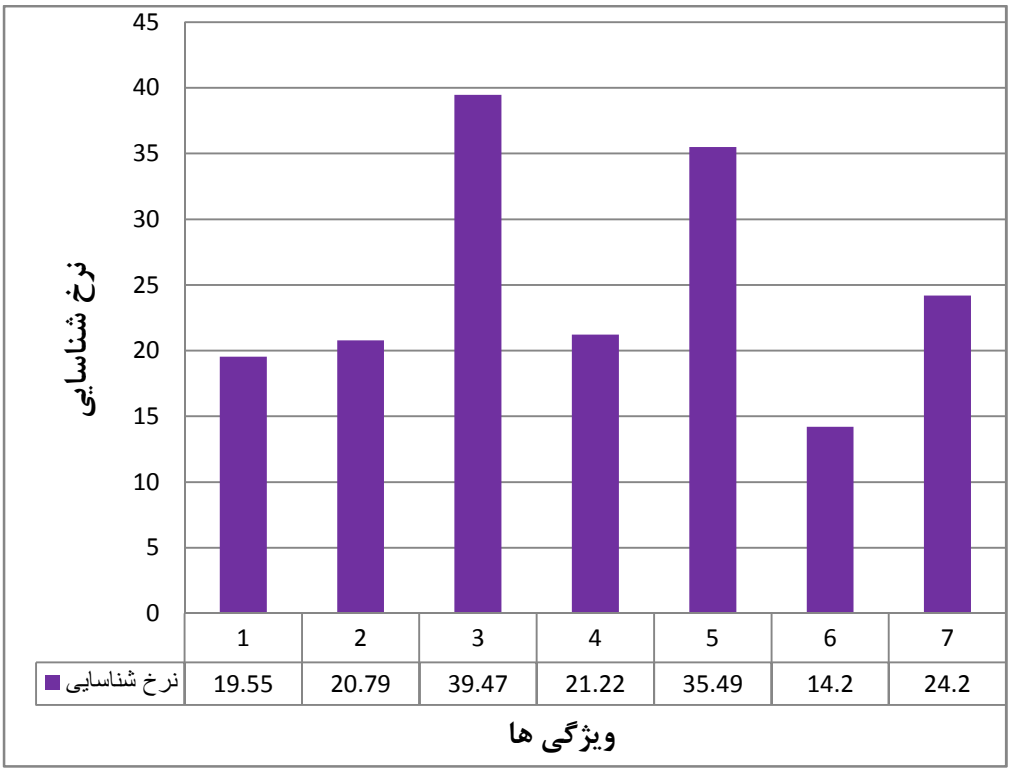
۴-۷-۳ کاهش تعداد فریم‌ها و کاهش سایز تصاویر با دستور ری‌سایز

این بار تعداد ۲۵ فریم از تصاویر را انتخاب و علاوه بر آن سایز تصاویر را با دستور ری‌سایز و روش درون‌یابی نزدیک‌ترین همسایگی[1] با اسکیل ۰.۵ کوچک‌تر نمودیم و ویژگی‌ها را از تصاویر سطح خاکستری محاسبه نمودیم. اما این بار بعد از استخراج ویژگی‌ها آن‌ها را به صورت سطری ذخیره نمودیم در نهایت هفت ویژگی شامل همبستگی بین دو فریم پشت سر هم و ضرایب DCT را پس از

[1] Nearest-neighbor Interpolation

یک اسکن زیگزاگ برای هرفریم ۱۰۰۰ ضریب و ضرایب DWT و ضرایب MFCC از هر کدام از ویژگی‌های قبل به جز همبستگی‌ها و نیز از خود تصویر را با این شرایط محاسبه نمودیم. در زیر نتایج حاصل از این ویژگی‌ها که به صورت زیر برچسب خورده‌اند آورده شده است.

۱-ضرایب همبستگی ۲- ضرایب DCT ۳- ضرایب DWT ۴- ضرایب MFCC از DCT ۵- ضرایب MFCC از DWT ۶- ضرایب MFCC از زیگزاگ DCT ۷- ضرایب MFCC از تصاویر

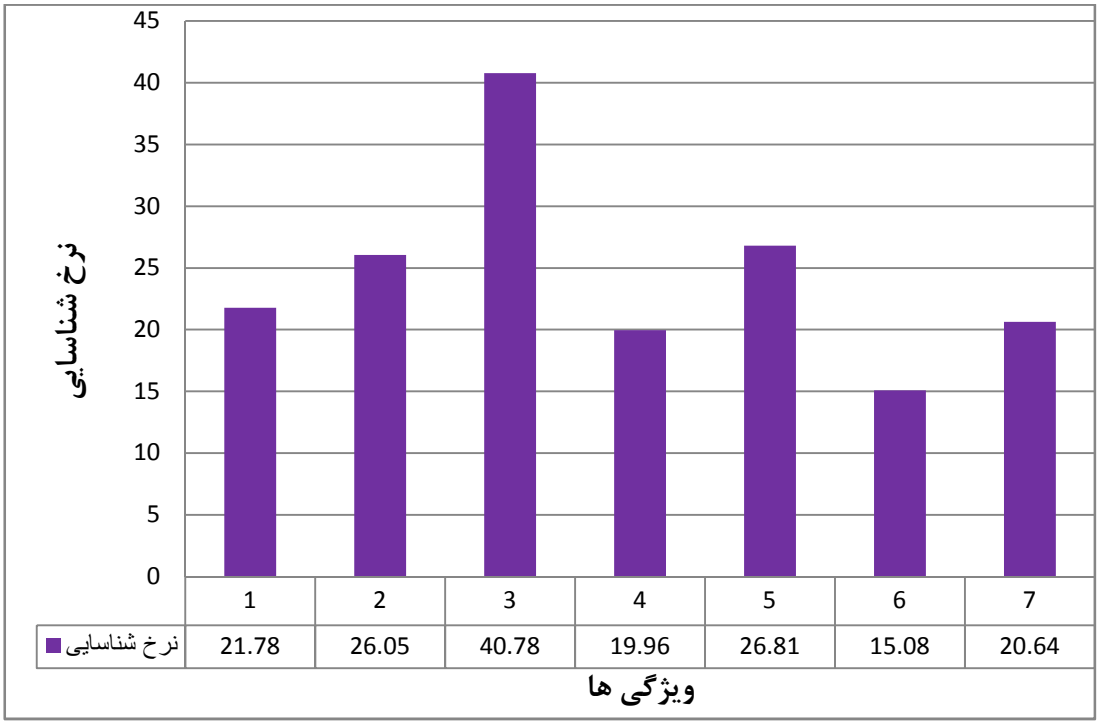


شکل ۴- ۱۴ نتایج حاصل از تصاویر کوچک شده با مقیاس ۰.۵و تعداد ۲۵ فریم

همین کار را دوباره با اسکیل ۰.۷ انجام دادیم و ویژگی‌ها را این بار به‌صورت ستونی ذخیره نمودیم و نتایج به‌صورت زیر به دست آمد.

این نتایج ماکزیمم درصد شناسایی بعد از چندین بار آموزش و تست بوده است. که از ۱۷۹ ویدیو برای تست و از ۲۰ ویدیو برای اعتبار سنجی و از ۳۸۱ ویدیو برای آموزش استفاده نمودیم و به شبکه

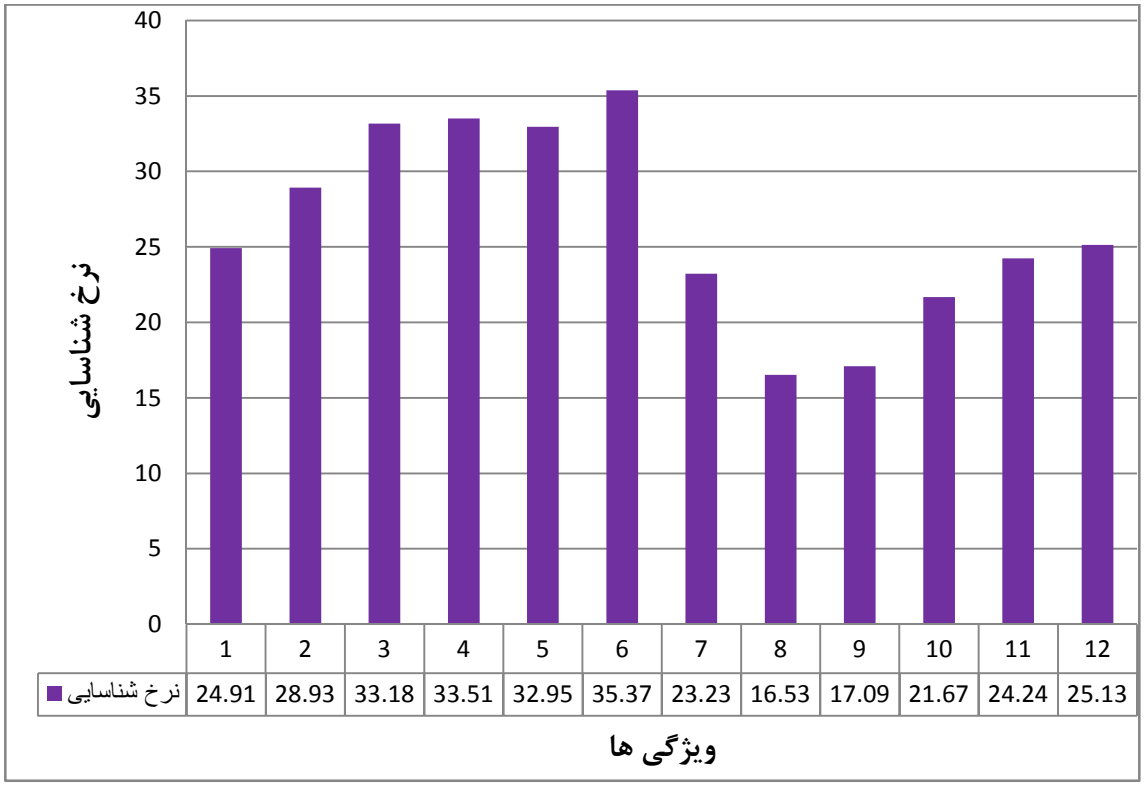
Feed-Forward با آموزش گرادیان نزولی با نرخ آموزش متغیر اعمال کردیم. نتایج حاصل در جدول زیر بیان شده است.



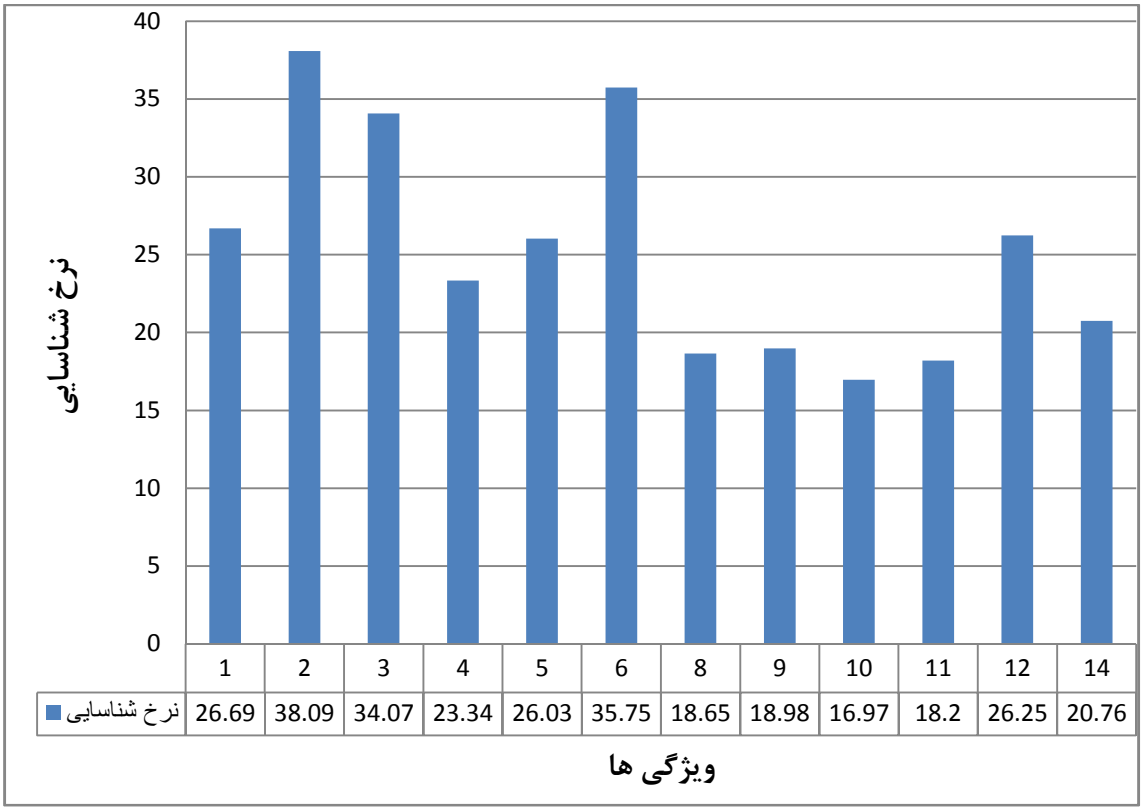
شکل ۴- ۱۵ نتایج حاصل از تصاویر کوچک شده با مقیاس ۰.۷و تعداد ۲۵ فریم

در مرحله‌ای دیگر با دو اسکیل ۰.۵ و ۰.۷ تصاویر را کوچک و ضرایب مختلف را محاسبه می‌کنیم با این تفاوت که در این جا از کل ضرایب DCT تعداد ۱۰۰۰و ۵۰۰ و ۱۰۰ و ۵۰ و ۱۰ را پس از اسکن انتخاب می‌کنیم. و از ضرایب DWT کل ضرایب حاصل از تبدیل را در نظر می‌گیریم. ضرایب مل فرکانسی را از ضرایب DCT و نیز یک دوم و یک چهارم و یک هشتم ضرایب پس از اسکن و از ضرایب DWT و همین ضرایب پس از اسکن و از خود تصویر محاسبه و به عنوان ورودی به شبکه اعمال می‌کنیم. نتایج حاصل از هر کدام از ویژگی‌ها و متوسط دقت شناسایی بعد از ۵ بار آموزش و تست بیان شده است. ویژگی‌ها به صورت زیر برچسب خورده‌اند.

۱-۱۰ضریب DCT پس از اسکن زیگزاگ، ۲-۵۰ ضریب DCT پس از اسکن زیگزاگ، ۳-۱۰۰ ضریب DCT پس از اسکن زیگزاگ، ۴- ۵۰۰ ضریب DCT پس از اسکن زیگزاگ، ۵- ۱۰۰۰ ضریب DCT پس از اسکن زیگزاگ، ۶ - کل ضرایب DWT، ۷- ضرایب MFCC از ماتریس DCT، ۸- ضرایب MFCC از ۱/۲ ضرایب DCT پس از اسکن زیگزاگ، ۹- ضرایب MFCC از ۱/۴ ضرایب DCT پس از اسکن زیگزاگ، ۱۰- ضرایب MFCC از ۱/۸ ضرایب DCT پس از اسکن زیگزاگ، ۱۱- ضرایب MFCC از ماتریس DWT، ۱۲- ضرایب MFCC از تصاویر.



شکل ۴- ۱۶ نتایج حاصل از ضرایب مختلف DCT با مقیاس ۰.۵



شکل ۴- ۱۷ نتایج حاصل از ضرایب مختلف DCT با مقیاس ۰.۷

مشاهده می‌کنیم که برای ضرایب DCT پس از اینکه با اسکن زیگزاگ به بردار تبدیل شد انتخاب بین ۵۰ تا ۱۰۰ ضریب از ضرایب اول بردار نتایج بهتری نسبت به سایر انتخاب‌ها دارد. ما با استفاده از کاهش سایز تصاویر و کاهش تعداد فریم‌ها نتوانستیم به نتیجه مطلوبی دست پیدا کنیم بنابراین کاهش ابعاد ویژگی‌ها با انجام این تغییرات بر روی تصاویر موثر نخواهد بود. استفاده از روش‌های کاهش ویژگی نتایج بهتری خواهد داشت.

۸-۴ نتیجه‌گیری

در این تحقیق پس از بررسی کارهایی که در زمینه شناسایی دیداری گفتار انجام شده، روشی بر مبنای ویژگی‌های فرکانسی – زمانی برای استخراج ویژگی از فریم‌های ویدیو ارائه گردید و نقش هر یک از این ویژگی‌ها در تشخیص مصوت‌ها بررسی شد. با توجه به اینکه این مصوت‌ها مربوط به کلمات تک سیلابی بودند و به صورت ویدیو بودند ابتدا فریم‌های مربوط به ادای این مصوت‌ها را از بقیه فریم-های ویدیو جدا نموده و بعد از یافتن ناحیه‌ای مطلوب پیرامون دهان برای کاهش ابعاد ویژگی‌ها، اقدام به استخراج ویژگی‌هایی همچون DCT و DWT و MFCC از تصاویر نمودیم. با توجه به اینکه کوچک نمودن سایز تصاویر نیز تاثیر چندانی در بهبود نتایج نداشت و نتوانست عملکرد ویژگی‌ها را افزایش دهد، تصمیم به استفاده از روشی برای کاهش ابعاد زیاد ویژگی‌ها و عملکرد بهتر شبکه گرفتیم. به همین دلیل از روش کاهش ویژگی LSDA استفاده نمودیم چون این روش در زمینه لب-خوانی به کار گرفته شده و در مقایسه با سایر روش‌های کاهش ویژگی در این زمینه نتایج بهتری به همراه داشته است. با استفاده از این روش سایز ویژگی‌ها را برای تمام ویژگی‌های استخراجی از فریم های تصویر به ۲۵ کاهش دادیم. مشخص گردید که استفاده از این روش کاهش ویژگی و اعمال آن به ویژگی‌های استخراج شده عملکرد آن‌ها را بهبود می‌بخشد. پس در این تحقیق بعد از اینکه هر فریم تصویر به سیگنال ۱ بعدی تبدیل شد، ضرایب MFCC و سایر ضرایب همچون ضرایب کسینوسی و ویولت و نیز ضرایب MFCC از ضرایب کسینوسی و ضرایب ویولت استخراج و عملکرد آن‌ها به همراه LSDA بررسی شد. در نهایت ضرایب MFCC که از ۱/۲ و ۱/۴ و ۱/۸ اولین ضرایب برداری که از اسکن زیگزاگ ماتریس ضرایب DCT به دست آمده بودند بهترین عملکرد را در پی داشتند. برای شناسایی از شبکه عصبی استفاده شد این شبکه دارای ۲۰ نرون میانی و یک خروجی بود و برای آموزش شبکه از روش گرادیان نزولی با نرخ آموزش متغیر استفاده گردید. از ۱۷۹ ویدیو برای تست و از ۲۰ ویدیو برای اعتبار سنجی و از ۳۸۱ ویدیو برای آموزش استفاده نمودیم نتایج با متوسط‌گیری

پس از ۵ بار آموزش و تست حاصل شد و بیشترین نرخ شناسایی ۹۵.۷۵٪ به دست آمد. می‌توان گفت که با استفاده از روش‌های مبتنی بر تصویر در مقایسه با روش‌های مبتنی بر مدل نتایج بهتری حاصل می‌گردد. ضرایب کسینوسی و ضرایب ویولت برای تشکیل بردارهای ویژگی بسیار مناسب هستند. اما با توجه به تعداد فریم‌ها در ویدیو استفاده از روش‌های کاهش ابعاد ویژگی لازم است. ضرایب MFCC نیز برای استخراج ویژگی از تصاویر مناسب هستند و می‌توانند برای شناسایی دیداری استفاده شوند. استفاده از اسکن زیگزاگ برای ماتریس ضرایب کسینوسی نیز می‌تواند به یافتن مهمترین ضرایب کسینوسی تصویر کمک نماید و دقت شناسایی را افزایش دهد. بنابراین استفاده از ضرایب کسینوسی پس از اسکن زیگزاگ برای استخراج ضرایب مل فرکانسی به بهبود و افزایش دقت سیستم شناسایی کمک خواهد کرد.

۹-۴ پیشنهاد ادامه کار

با توجه به اینکه در این رابطه کارهای بسیاری در زبان فارسی مشاهده نشده است. موارد پیشنهادی برای ادامه کار به شرح زیر می‌باشد.

۱- تهیه پایگاه داده جامعی، با رعایت فاصله و شرایط نوری و نیز تنوع و کثرت گویندگان

۲- بهبود روش‌ها به منظور تشخیص صحیح در سایر شرایط روشنایی

۳- استفاده از گفتار برای بهبود نتایج

۴- استفاده از سایر روش‌های استخراج ویژگی به کار رفته در پردازش گفتار همچون ضرایب PLP و Root MFCC و .....

۵- بسط مسئله به تشخیص صامت‌های فارسی و یا تشخیص کلمه به جای تشخیص مصوت‌ها

مراجع

[1] T Chen, "Audiovisual speech processing". IEEE Signal Processing Magazine , Vol.18(1), pp: 9–21, (2001).

[2] صادقی، وحیده السادات، "تشخیص مصوت در کلمات تک سیلابی و دو سیلابی فارسی،" پایان نامه کارشناسی ارشد، دانشگاه سمنان، ۱۳۸۵

[3] E.D.Petajan, "Automatic Lipreading to Enhance Speech Recognition," PhD thesis, University of Illinois at Urbana-Champain, 1984.

[4] M. Kass, A.Witkin, and Terzopoulos, " Snakes: Active Contour Models," *International Journal of Computer Vision* , pp.321-331,1988.

[5] C. Bregler and Y. Konig, " Eigenlips For Robust Speech Recognition," *in Proc. IEEE Conf. Acoustics, Speech and Signal Processing*, pp.669-672, 1994.

[6] Takeshi Saitoh and Ryosuke Konishi , " Word Recognition based on Two Dimensional Lip Motion Trajectory, " *international Symposium on Intelligent Signal Processing and Communication System(ISPACS2006)* ,pp.287-290. 12-15 Dec, 2006

[7] میر هادی سید عربی، علی آقا گلزاده، سهراب خان محمدی، "تعقیب اتوماتیک حرکات لب و نقاط ویژه آن با استفاده از کانتور فعال"، *چهاردهمین کنفرانس مهندسی برق ایران* ۲۰۰۶ ICEE.

[8] T.F. Cootes , C.J. Taylor, D.H. Cooper, and J. Graham, “Active Shape Models-Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38-59, Jan. 1995

[9] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal .Mach. Intell*., vol. 24, no. 2, .pp. 198–213, Feb. 2002

[10] Juergen Luettin,Neil A. Thacker ," Speechreading using probabilistic Models," *Computer Vision and Image Understanding*, Vol.65,No.2, pp.163-178, February 1997

[11] S.L.Wang , W.H.Lau , S.H.Leung, et al. " A real-time automatic lipreading system,"

International Symposium on Circuits and Systems, No.2, pp.101-104,IEEE, Vancouver , Canada, May 2004.

[12] D. Thambiratnam , T. Wark , S.Sridharan and V.Chandran , "Speech Recognition in Adverse Environments using Lip Information," *Speech and Image Technologies for Computing and Telecommunications, IEEE TENCON 1997*, Vol.1, pp.149-152, 4Dec,1997

[13] Tanveer A Faruquie, Abhik Majumdar, Nitendra Rajput, L V Subramaniam,"Large Vocabulary Audio-Visual Speech Recognition Using Active Shape Models," *Pattern Recognition ,2000,15th International Conference,* Vol.3, pp.106-109,2000.

[14] A.L.Liew, et al," Lip contour extraction from color images using a deformable model," *The Journal of the Pattern Recognition Society*, No.35, 2949-2962, 2002

[15] Stefan Horbelt, Jean-Luc Dugelay ," Active Contours For Lipreading Combinning With Templates," *15th GRETST Symposium on Signal and Image processing*,pp.18-22, September 1995,france.

[16] Mohammad Mehdi Hosseini, Abdorreza Alavi Gharahbagh and Sedigheh Ghofrani ," Vowel Recognition by Using the Combination of Haar Wavelet and Neural Network," *KES'10 Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems*, Part I,pp.331-339, 2010.

[17] M.M,Hosseini, S.Ghofrani ," Automatic Lip Extraction Baced On Wavelet Transform," *IEEE GCIS*, pp.393-396, 2009,China.

[18] Dahai Yu, Ovidiu Ghita, Alistair Sutherland, Paul F. Whelan," A PCA based Manifold Representation for Visual Speech Recognition,"In: CIICT 2007, Proceedings of the China-Ireland International Conference on Information and Communication Technologies, 28-29 August 2007, Dublin, Ireland.

[19] Y. L. Tian and T. Kanade," Robust Lip Tracking by Combining Shape, Colour and Motion," *Proc. of the Asian Conference on Computer Vision*, pp.1040-1045, 2000.

[20] Kim YongMin, Li Hong Zuo, " A Lip Reading Method Based on 3-D DCT and 3-D HMM," *International Conference on Electronics and Optoelectronics*, vol.1,pp.115-119, IEEE 2011.

[21] H. Ertan Cetingul, Yucel Yemez, Engin Erzin and A. Murat Tekalp," Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading," *IEEE Transactions on Image Processing*, VOL. 15, NO. 10, October 2006.

[22] Xiaoping WANG, Yufeng HAO, Degang FU, Chunwei YUAN, "ROI Processing for Visual Features Extraction in Lip-reading"**,** *IEEE Int. Conference Neural Networks & Signal Processing*, pp. 178-181, 7-11 June 2008.

[23] Liang Yaling, Yao Wenjuan, Du Minghui, "Feature Extraction Based on LSDA for Lipreading", IEEE 2010.

[24] I. Shdaifat and R. Grigat,D. Langmann," A System for Automatic Lip Reading ," *International Conference on Audio-Visual speech Processing*,4-7September , 2003.

[25] Amin Banitalebi, Maryam Moosaei, Gholam Ali Hossein zadeh ," An Investigation on the usage of Image Quality Assessment in visual speech Recognition," *The 6th Iranian machine vision & image processing conference* , 27-28 October 2010.

[26] Z. Wang and E.P. Simoncelli, " Translation insensitive image similarity in complex wavelet domain," *in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp.573-576, , Mar. 2005

[27] Vahideh Sadat Sadeghi, Khashayar Yaghmaie," vowel recognition using neural network," *IJCSNS International, Journal of Computer Science and Network Security*, VOL.6 No.12, December 2006.

[28] S.L.Wang, A.W.C.Liew, W.H.Lau,and S.H.Leung ," An Automatic Lipreading System for Spoken Digits With Limited Training Data," *IEEE Transactions on Circuits and Systems for Video Technology*, VOL. 18, NO. 12, December 2008.

[29] N. Eveno, A. Caplier, P.Y. Coulon, New color transformation for lips segmentation, in: *Proceedings of IEEE Fourth Workshop on Multimedia Signal Processing*, pp. 3–8, Cannes, France, October 2001.

[30] Wark,T.,sridharan,S.,and Chaandran,V."An approach to statistical lip modelling for speaker identification via chromatic feature extraction" .*In proceeding of the IEEE International conference on Pattern Recognition*, Vol.1, pp 123-125, Aug 1998.

[31] Coianiz,T.,Torresani,L.,and Caprile,B."2D deformable models for visual speech analysis".In [Stork and Hennecke,1996] , pp 391-398.

[32] Vogt, M. "Fast matching of a dynamic lip model to color video sequences under regular illumination conditions".In[Stork and Hennecke,1996], pp.399-407.

[33] Hamed Talea, Khashayar Yaghmaie,"Automatic visual speech segmentation", 3rd International Conference on Communication Software and Networks, pp.4854-4858, 2011 IEEE

[34] F. G. Hashad, T. M. Halim S. M. Diab, and B. M. Sallam," A New Approach for Fingerprint Recognition Based on Mel Frequency Cepstral Coefficients", *International Conference on Computer Engineering & System*, pp. 263-268, 14-16 Dec, 2009.

[35] Shikha Gupta1, Jafreezal Jaafar, Wan Fatimah wan Ahmad3 and Arpit Bansal, " Feature Extraction Using Mfcc" , *Signal & Image Processing : An International Journal (SIPIJ)* Vol.4, No.4, August 2013

[36] M. M. M. Fahmy, " Palmprint recognition based on Mel frequency Cepstral coefficients feature extraction", *Ain Shams Engineering Journal*, p. 9, 2010.

[37] N. Puviarasan , S. Palanivel ,"Lip reading of hearing impaired persons using HMM", 2010 Elsevier Ltd, Expert Systems with Applications 38 (2011).pp. 4477–4481,

[38] Md. Rashidul Hasan, Mustafa Jamil Md. Golam Rabbani,Md. Saifur Rahman, "Speaker Identification using Mel Frequency Cepstral Coefficients", *3rd International conference on Electrical and computer engineering ICECE* 2004,Dec 2004.

[39] T. M. Talal and A. El-Sayad, "Identification of Satellite Images Based on Mel Frequency Cepstral Cofficients, pp.274-282, IEEE 2009.

[40] Sangeeta Biswas" MFCC based Face Identification" Titech Japan, 2009.

[41] Deng Cai, Xiaofei He, Kun Zhou, "Locality Sensitive DiscriminantAnalysis," International Joint Conference on Artificial Itelligence. Hyderabad: morgan Kaufmann Publishers 2007. pp.708-713.

Abstract

Visual features have been widely used to improve the performance of speech recognition. In this thesis time - frequency features extracted from the images of the speaker 's mouth and extracted features are used as input parameters to a neural network system for recognition. Because we used the video images so we got to work a different number of video frames. First separated the frames manually and then selected the area around the mouth and desired features for the area of each frame obtained. To improve performance and reduce the dimensions of features, we used dimensionality reduction technique LSDA. Using this approach we have reduced the size of our feature. The database consists of different individuals, that have been uttered monosyllabic words 2 or 3 times. Finally the vowel recognition rate 95.75 was achieved.

Keyword:

Lip reading,Vowel recognition, Time-frequency features, Feature dimension reduction, Neural networks

Shahrood University of Technology

Faculty of Electrical and Robotic

Time-frequency feature extraction for visual recognition of persian vowels

Nasrin Yadegar Khosravieh

Supervisor(s): Dr Hossien Marvi

Date : February 2014