بازشناسی متن فارسی قدیمی با استفاده از توصیفگرهای ویولت در روش منهتن

مائده رضائی فر[۱]، افشین ابراهیمی[۲]
[۱]دانشجو دانشگاه صنعتی سهند تبریز،دانشکده برق ، m_rezaeifar@sut.ac.ir
[۲]دانشیار دانشگاه صنعتی سهند تبریز، دانشکده برق ، a_ebrahimi@sut.ac.ir

چکیده - در این مقاله به ارائه روش جدیدی برای بازشناسی اسناد چاپی قدیمی فارسی مبتنی بر شکل کلی کلمات با استفاده از بسته موجک و تبدیل بسته موجک پرداخته‌ایم. بنابه ویژگی‌های خاص زبان فارسی و عربی بازشناسی این زبان‌ها بر اساس شکل کلی کلمات نسبت به زبان‌های لاتین و چینی متفاوت‌تر خواهد بود. کارهای انجام شده، بیشتر بر متون دست نویس متمرکز بوده است در حالی‌که روش ارائه شده در این مقاله برای بازشناسی یک کتاب چاپی قدیمی فارسی است که داری فونت خاص و چاپ سنگی است. روش ارائه شده بر پایه شکل کلی کلمات و استفاده از روش منهتن روی توصیفگرهای موجک می‌باشد. با تهیه‌ی واژه‌نامه مناسب از کتاب قدیمی فارسی که از ۳۷۸۶۱ زیر کلمه مشابه و غیر مشابه تشکیل شده است و استفاده از کد نقاط و فشرده سازی بردارهای ویژگی، نرخ بازشناسی مناسبی با مقدار ۸۸٪ ارائه شده است.

کلید واژه- بازشناسی متن قدیمی، بسته تبدیل موجک، تبدیل موجک، روش منهتن، نویسه خوان نوری.

۱- مقدمه

از کاربردهای بازشناسی الگوهای تصویری، نویسه خوان نوری حروف می‌باشد. بازشناسی متن، یکی از زمینه هایی است که در چند دهه اخیر، توجه زیادی را به خود اختصاص داده است[۴-۱]. یک سیستم نویسه خوان نوری، تصویر اسناد را به اسنادی به صورت کدهای کاراکتری تبدیل می‌کند. این عمل در کاهش حجم اسناد تأثیر بسیار زیادی دارد و همچنین در ذخیره سازی اسناد در تعداد زیاد و انتقال این اسناد بوسیله شبکه‌های کامپیوتری بسیار مفید می‌باشد. یک سیستم نویسه خوان نوری دارای مراحل مختلفی بوده و کارایی این سیستم به دقت و عملکرد هر یک از این مراحل بستگی دارد. این مراحل شامل روبش سند، پیش پردازش، استخراج ویژگی، تشخیص کاراکترها و پس پردازش می‌باشد. سیستم‌های بازشناسی متن براساس داده ورودی دریافتی به شاخه های متن چاپی و متن دست نویس تقسیم می‌شوند. اولین مرحله در بازشناسی متن، دریافت داده و تبدیل آن به شکل دیجیتالی قابل استفاده در کامپیوتر است. بازشناسی متن دستنویس به دو روش برخط و برون خط است. سیستم بازشناسی هر یک متفاوت می‌باشد. ورودی سیستم بازشناسی برون خط، تصویر روبش شده[۶] متن دستنویس و اسناد چاپی است [۵]. روش‌های بازشناسی متن فارسی، به دو دسته کلی تقسیم می‌شوند. بازشناسی بر اساس شکل کلی کلمات و بازشناسی بر پایه جداسازی. در روش اول، کل زیر کلمه به عنوان یک شکل در نظر گرفته شده و سپس تشخیص داده می‌شود و در روش دوم، ابتدا زیر کلمه ورودی به حروف و زیر حروف تجزیه شده و سپس بازشناسی صورت می‌گیرد. در برخی از موارد از ترکیب این دو روش نیز استفاده می‌شود. در این مقاله، ما با استفاده از روش مبتنی بر شکل کلی کلمات و همچنین استفاده از دانش‌های جانبی مانند نقاط، علائم و برخی قواعد دستوری ساده[۶] زبان فارسی که در ادامه به توضیح آن می‌پردازیم، به بازشناسی این کتاب چاپی قدیمی پرداخته ایم.

کارایی یک سیستم نویسه خوان نوری(OCR) وابسته به این است که ویژگی هایی که هر الگو را نمایش می‌دهند، چگونه تعریف می‌شوند. روش‌های متعددی برای بازشناسی متون قدیمی در اسناد لاتین و چینی پیشنهاد شده است. ولی فضای تحقیق در اسناد چاپی قدیمی فارسی و همچنین در اسناد عربی مورد نیاز است. علاوه بر این، در ۵۰ سال گذشته بیشتر تلاش‌ها در بازشناسی متون چاپی عربی صورت گرفته است و محک آن‌ها بر روی دیتابیس‌های گذشته بوده است[۶]. تلاش بر این بوده است که با روش بکار رفته در این مقاله نرخ بازشناسی یک متن قدیمی را افزایش دهیم.

۲- کاراکترهای زبان فارسی:

نحوه نگارش زبان فارسی نسبت به زبان‌های دیگر از تفاوت عمده‌ای برخوردار است. این تفاوت در بازشناسی متون فارسی مشکلات مخصوص به خود را دارد. حالت‌های مختلف حروف الفبای فارسی با توجه به موقعیت مکانی آن‌ها در کلمات در

جدول (۱) حالت‌های مختلف حروف الفبای فارسی نشان داده شده است.

جدول (۱) حالت‌های مختلف حروف الفبای فارسی[۷]

Isolated	Initial	Medial	Final	Roman	Name	Isolated	Initial	Medial	Final	Roman	Name
ا	ا	ـا	ـا	á	alef	ص	صـ	ـصـ	ـص	ṣ	sád
ب	بـ	ـبـ	ـب	b	be	ض	ضـ	ـضـ	ـض	ḍ	zád
پ	پـ	ـپـ	ـپ	p	pe	ط	طـ	ـطـ	ـط	ṭ	tá
ت	تـ	ـتـ	ـت	t	te	ظ	ظـ	ـظـ	ـظ	ẓ	zá
ث	ثـ	ـثـ	ـث	th	se	ع	عـ	ـعـ	ـع	ʿ	ayn
ج	جـ	ـجـ	ـج	j	jim	غ	غـ	ـغـ	ـغ	gh	ghayn
چ	چـ	ـچـ	ـچ	ch	che	ف	فـ	ـفـ	ـف	f	fe
ح	حـ	ـحـ	ـح	ḥ	ḥe	ق	قـ	ـقـ	ـق	q	qáf
خ	خـ	ـخـ	ـخ	kh	khe	ک	کـ	ـکـ	ـک	k	káf
د	د	ـد	ـد	d	dál	گ	گـ	ـگـ	ـگ	g	gáf
ذ	ذ	ـذ	ـذ	dh	zál	ل	لـ	ـلـ	ـل	l	lám
ر	ر	ـر	ـر	r	re	م	مـ	ـمـ	ـم	m	mím
ز	ز	ـز	ـز	z	ze	ن	نـ	ـنـ	ـن	n	nún
ژ	ژ	ـژ	ـژ	zh	zhe	و	و	ـو	ـو	v/ú	váv
س	سـ	ـسـ	ـس	s	sin	ه	هـ	ـهـ	ـه	h	he
ش	شـ	ـشـ	ـش	sh	shin	ی	یـ	ـیـ	ـی	y/í	ye

از مشکلات کد نقاط می‌توان به اتصال چسبندگی نقطه‌های حروف به یکدیگر و یا به بدنهٔ خود زیر کلمه اشاره کرد. این مشکلات در تصاویر نویزدار سبب ایجاد مساحت بزرگ‌تری خواهد شد به‌طوری‌که به اندازه یک پیکسل می‌رسد و با حذف نویز، آن نقطه و یا نقاط نیز حذف می‌شوند. به همین دلیل با روش‌های مانند مورفولوژی به نازک سازی حروف به بدنه و یا نازک سازی دو زیر کلمه به یکدیگر پرداخته ایم.

وجود سرکش در بعضی از حروف مانند "ک" و "گ" ، علامت مد روی حرف الف، همین‌طور حروف دسته‌داری مانند "ط" و "ظ" تفاوت در شکل‌های بعضی حروف در زبان فارسی را نشان می‌دهند. همچنین تفاوت شکل در برخی حروف دیگر علاوه بر تعداد در محل قرار گرفتن نقاط آن‌ها نیز هست مانند {"ب"، "پ"، "ت"، "ث"}.

یکی از خصوصیات متمایز کننده در متون فارسی خط زمینه و یا خط مبنا است. خط زمینه به خطی گفته می‌شود که بیشترین تعداد نقاط متن (پیکسل‌های سیاه) را دارا است و قسمت پایینی بدنه‌ی کاراکترها را در یک خط متن دنبال می‌کند[۸].

۳= واژه‌نامه:

در این پروژه از تصاویر روبش شده یک کتاب چاپی قدیمی به‌عنوان پایگاه داده الکترونیکی استفاده کرده‌ایم و هدف ما بازخوانی متن و تبدیل آن به متن قابل ویرایش در Word است.صفحات کتاب را با دقت dpi۶۰۰ در کتابخانه مرکزی اسکن کرده‌ایم. به دلیل قدیمی بودن متن و نیز اسکن این صفحات، وجود نویز، گسستگی و یا چسبندگی کلمات و نقاط به یکدیگر و یا به بدنه‌ی زیر کلمات را خواهیم داشت. برای ایجاد واژه‌نامه ابتدا تمام زیر کلمات تصاویر را جدا کرده و در یک ساختار با فیلد image ذخیره کرده‌ایم و تمام زیر کلمات مشابه و غیرمشابه را جدا نموده‌ایم. برای این کار طبق برنامه نوشته شده معادل متنی هر زیر کلمه به برنامه داده می‌شود اگر در واژه نامه موجود بود، محل آن نمایش داده می‌شود در غیر اینصورت از آخر در پایگاه داده تهیه شده، ذخیره می‌گردد. با این روش توانستیم از مجموع ۳۷۸۶۱ زیر کلمه که همه دارای یک قلم و اندازه ثابت بوده‌اند، تعداد ۱۶۶۹ زیرکلمه غیرمشابه ایجاد نموده و به‌عنوان واژه‌نامه ذخیره نماییم.

۴- تبدیل موجک و بسته تبدیل موجک:

۴-۱ تبدیل موجک:

در سال‌های اخیر تبدیل‌های موجک در پردازش تصویر بر بسیاری از تکنیک‌های پیشرفته چیره شده‌ان یک سطح تبدیل موجک دو بعدی می‌تواند روی هر تصویر f به اندازه $M \times N$ تعریف شود[۹].

در رویکرد مبتنی بر شکل کلی کلمات یا زیر کلمات، توصیفگرها به طور مستقیم از تصویر کلمه یا زیر کلمه استخراج می‌شوند و با استفاده از یک واژه نامه تصویری، کلمه مناسب استخراج می‌شود. در این تحقیق، ویژگی‌های شکل زیر کلمات با استفاده از تبدیل موجک استخراج شده است و بازشناسی زیر کلمات با استفاده از این توصیفگرهای موجک انجام گردیده است. نتایج خوب بدست آمده از تبدیل موجک در پردازش تصویر سبب شده است تا با بکارگیری از محلی کردن و تمرکز در نقطه‌ی بخصوص مکانی و جزئیات محلی سازی فرکانسی، تبدیل موجک به ابزار موثری برای استخراج ویژگی و بازشناسی متون تبدیل شود. در بازشناسی کلمات و اعداد موجود در متن به زبان های مختلف، تبدیل موجک به روش‌های مختلف برای استخراج

ویژگی و بازشناسی آنها استفاده شده است. در هریک از این روش ها تبدیل موجک مادر استفاده شده نیز متفاوت است. انواع تبدیل موجک مورد استفاده در بازشناسی کاراکتر در کارهای پیشین را می‌توان به : تبدیل موجک گسسته دو بعدی، تبدیل موجک گسسته مختلط، تبدیل موجک پیوسته جهت دار، موجک چندگانه متعامد، گشتاورهای موجک، موجک بسته ای، موجک بسته‌ای M-باند است که ساده ترین راه برای محاسبه‌ی یک تبدیل موجک گسسته (DWT) دو بعدی یک تصویر، بکار بردن یک تبدیل یک بعدی روی سطرها و ستون های تصویر بطور مجزا و سپس انجام نمونه برداری کاهشی است.

۴-۲ بسته تبدیل موجک:

روش تبدیل بسته موجک یک تعمیم تجزیه‌ی موجک است که یک آنالیز سیگنال غنی‌تر را پیشنهاد می‌دهد. در روندهای تجزیه‌ی موجک متعامد دو بُعدی، مرحله‌ی کلی(جامع) ضرائب تخمین را به چهار زیر بخش تقسیم می‌کند. بعد از تقسیم‌بندی، یک زیر تصویر از ضرائب نرم(پایین گذر) و سه زیر تصویر متناظر با ضرائب جزئیات در جهت‌های عمودی و افقی و مورب بدست می‌آید. این تقسیم‌بندی روی زیر تصویر پایین گذر دنبال می‌شود در حالیکه جزئیات متوالی دوباره آنالیز نمی‌شوند[۹].

در موقعیت بسته‌ی موجک متناظر، هر زیر تصویر ضرائب جزئیات به چهار بخش با استفاده از همان روش مانند تقسیم‌بندی زیر تصویر پایین گذر تجزیه می‌شود[۹] شکل (۱) ساختار تجزیه سطح دوم تبدیل بسته موجک است . تبدیل بسته موجک گسسته‌ی دوبُعدی برای یک تصویر گسسته‌ی $N \times M$ بُعدی به نام A، بالای سطح $(p \leq min(log_2 N, log_2 M))p + 1$ به طور بازگشتی بر حسب ضرائب در سطح p به صورت زیر تعریف می‌شود[۱۰]

$$C^{p+1}_{4k,(i,j)} = \sum_m \sum_n h(m)h(n)\, C^{p}_{k,(m+2i,n+2j)}$$
$$C^{p+1}_{4k+1,(i,j)} = \sum_m \sum_n h(m)g(n)\, C^{p}_{k,(m+2i,n+2j)}$$
$$C^{p}_{4k+2,(i,j)} = \sum_m \sum_n g(m)h(n)\, C^{p}_{k,(m+2i,n+2j)}$$
$$C^{p+1}_{4k+3,(i,j)} = \sum_m \sum_n g(m)g(n)\, C^{p}_{k,(m+2i,n+2j)}$$

که $C^{\cdot}_{\cdot,(i,j)}$ تصویر A است. در هر سطح، تصویر C^{p}_{k} به چهار زیر تصویر به اندازه‌ی یک چهارم تصویر C^{p}_{k}، به نام‌های C^{p+1}_{4k}، C^{p+1}_{4k+1}، C^{p+1}_{4k+2}، C^{p+1}_{4k+3} تجزیه می‌شود. فیلتر پایین گذر h و فیلتر بالاگذر g متناظر با توابع پایه موجک مختلف، متفاوت می‌باشد.

L1	H1	L2	H2
V1	D1	V2	D2
L3	H3	L4	H4
V3	D3	V4	D4

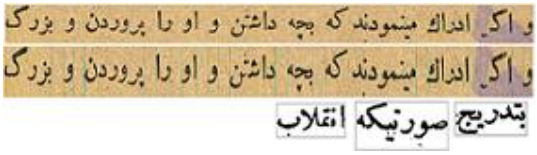
شکل (۲): ساختار تجزیه سطح دوم تبدیل بسته موجک[۱۱]

۵- الگوریتم پیشنهادی:

ابتدا تصویر ورودی را دریافت کرده و به جداسازی هر یک از خطوط آن پرداخته‌ایم. در کارهای گذشته یک ردیف به عنوان خط مبنا شناخته می‌شد، در حالیکه خط فارسی دارای عرض قلم می‌باشد به تعریف محدوده‌ی مناسب برای خط مبنا پرداخته‌ایم . از افکنش افقی برای پیداکردن نواحی سفید متن و با بسط دادن خط مبنا از بالا و پایین و تعیین یک آستانه مناسب محدوده خط را بدست آورده‌ایم و خطوط را از یکدیگر جدا نموده‌ایم[۱۲]. در شکل(۲) تصاویر تعیین خطوط از صفحات را نمایش داده شده است.

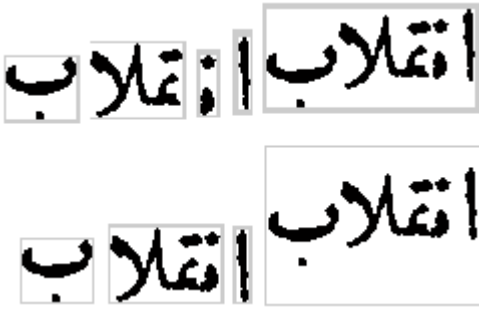
شکل(۲): تصاویر تعیین خطوط صفحات کتاب قدیمی

سپس به جداسازی کلمات و زیر کلمات هرخط پرداخته و با افکنش عمودی فواصل سفید بین ستون‌های کلمه را پیدا کرده و به جداسازی کلمات هر خط پرداخته‌ایم[۱۲]. شکل(۳) جداسازی کلمات و زیر کلمات کتاب قدیمی فارسی را نشان می-دهد.



شکل(۳): جداسازی کلمات و زیر کلمات کتاب قدیمی

به منظور انجام پیش پردازش‌های لازم به جهت از بین بردن همپوشانی و مشکل چسبندگی زیر کلمات به یکدیگر از تکنیک برچسب زنی استفاده نموده‌ایم[۱۳] همچنین با تعیین $thereshold \leq 2$ توانستیم گسستگی بین کلماتی که روی خط زمینه متمرکز هستند را برطرف نماییم شکل (۴) رفع گسستگی در کلمه با روش برچسب زنی را نشان می‌دهد.



شکل(۴): رفع گسستگی در زیر کلمه با تکنیک برچسب زنی

۵-۱ تجزیه تصویر زیرکلمه نرمالیزه شده با تبدیل موجک:

در الگوریتم تبدیل موجک هدف استفاده از ضرائب زیر باند تقریب سطح دوم حاصل از تجزیه موجک به عنوان ویژگی برای زیرکلمات پایگاه داده است. اندازه تصویر ورودی را به ابعاد۴۰ ×۴۰ نرمالیزه کرده‌ایم[۱۱]. تصویر ورودی نرمالیزه شده وارد فیلتر موجک هار می‌شود و تا دو سطح تجزیه می‌گردد. سپس ضرائب تقریب سطح دوم به عنوان ویژگی انتخاب می شوند، به این صورت که ضرائب این زیر باند دو بعدی به ابعاد ۵×۵ در یک بردار اندازه‌ی ۲۵×۱ ریخته می‌شوند و بردار ویژگی را تشکیل می دهد. جدول (۳) بردارهای ویژگی‌های مختلف تبدیلموجک را نمایش می‌دهد. از ویژگی‌های مختلف تنها ویژگی سیملت ۸، نیاز به کاهش بعد (PCA) دارد[۱۴].

جدول (۳): بردارهای ویژگی‌های مختلف تبدیل موجک

ویژگی/ویولت	Haar	Db1	Db2	Coif1	Sym8 (pca) شده
Haar	۱۶۶۹X۲۵	-	-	-	-
Db1	-	۱۶۶۹X۲۵	-	-	-
Db2	-	-	۱۶۶۹X۴۹	-	-
Coif1	-	-	-	۱۶۶۹X۸۱	-
Sym8	-	-	-	-	۱۶۶۹X۳۲۴/۳۸

۵-۲ تجزیه تصویر زیر کلمه نرمالیزه شده با بسته تبدیل موجک:

ضرائب زیر باند سطح[۱۵] به عنوان ویژگی به صورت یک بردار ۷۲۹ تایی در سیملت ۸ استفاده می‌شود. زیر باند به صورت یک ماتریس ۲۷ × ۲۷ تایی بدست می‌آید که همه‌ی سطرهای آن به ترتیب در یک بردار ستونی زیر هم اضافه می‌شوند و بردار ویژگی را به صورت یک بردار ۷۲۹ × ۱ تشکیل می‌دهند. اندازه تصویر ورودی در بسته تبدیل موجک به ابعاد ۶۴×۶۴ نرمالیزه شده‌اند.

۵-۳شبکه تشخیص نقاط:

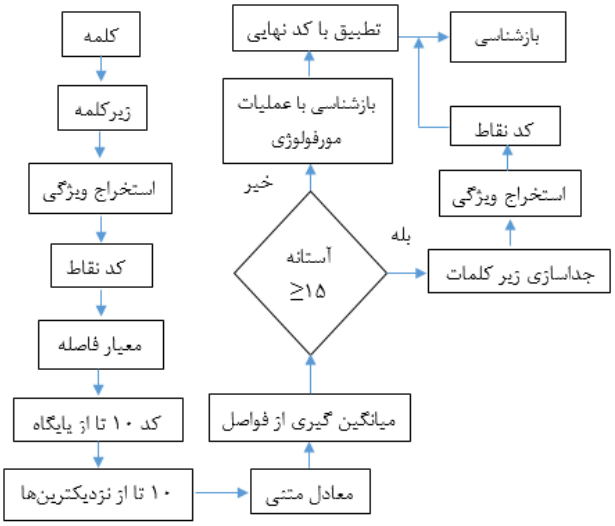
برای آموزش ویژگی‌های بکار رفته از شبکه عصبی دو لایه «FeedForward » با ۴۰ نورون در لایه خروجی و توابع انتقال «tansig» در لایه مخفی و «purelin» در لایه خروجی استفاده شده است که از تعداد ۵۰۰ تکرار و نرخ یادگیری ۳ استفاده نموده‌ایم. شبکه دارای ۵ کلاس ورودی و ۵ کلاس خروجی خواهد بود که کلاس ۱ برای تک نقطه، کلاس ۲ برای دونقطه، کلاس ۳ برای سه نقطه بالا، کلاس ۴ برای سه نقطه پایین و کلاس ۵ برای علامت مد الف خواهد بود .جدول (۴) نمونه زیر کلمات و کد نقاط آن‌ها را نشان داده شده است.

جدول (۴) نمونه زیر کلمات و کد نقاط آن‌ها

زیرکلمه	کدنقاط زیرکلمه	زیرکلمه	کدنقاط زیرکلمه
پیشین	Pysyn	پر	p
محیط	Y	سخیفه	nyn

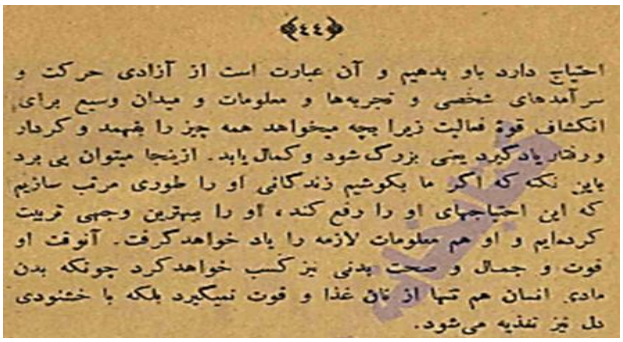
۵-۴ بازشناسی زیر کلمات:

فاصله ویژگی یک زیرکلمه از پایگاه داده با کل ویژگی زیرکلمات پایگاه داده محاسبه می‌شود و کمترین آن را به عنوان محل زیرکلمه بازشناسی شده انتخاب کرده‌ایم. یکبار از کد نقاط برای بازشناسی و تعیین محل زیرکلمات استفاده کرده‌ایم و در حالت دوم عملیات مورفولوژی را به آن اضافه نموده‌ایم و نتایج بدست آمده را مورد بررسی قرار داده‌ایم. فلوچارت روش پیشنهادی در شکل(۵) نشان داده شده است.



شکل(۵): فلوچارت روش منهتن

ابتدا کلمه ورودی به زیر کلمه تجزیه گردیده و سپس به استخراج ویژگی‌های بکار رفته پرداخته شده و کد نقاط آن‌ها بدست آمده است و از معیار فاصله منهتن استفاده شده است. کد ده زیر کلمه از پایگاه داده را نشان داده و ده عدد از نزدیکترین زیر کلمات مشابه به زیر کلمه مورد نظر را از نظر ساختار بدنی و کد نقاط نشان می‌دهد و در نهایت معادل متنی هرکدام از زیر کلمات را می‌توانیم ببینیم. سپس فاصله‌های این زیر کلمات بدست آمده و عملیات میانگین گیری و فاصله از میانگین ها را حساب کرده‌ایم. بزرگترین فاصله تعیین شده اگر از ۱۵کوچکتر بود به این معنی است که در زیر کلمه مورد نظر چسبندگی و یا گسستگی وجود دارد و با استفاده از عملگرهای انبساطی و انقباضی و حذف کشیدگی، به‌درستی زیر کلمه مورد نظر را بازشناسی می‌کنیم و اگر میانگین بزرگترین فاصله، از ۱۵ بزرگتر بود در اینصورت به بازشناسی زیر کلمه بدون عملیات مورفولوژی می‌پردازیم. تصاویر (۱۰)، (۱۱) و (۱۲) به ترتیب صفحات اصلی کتاب و بازشناسی شده را با ویژگی‌های db1 و wdb1 نشان می-دهد.



تصویر(۱۰): تصویر اصلی کتاب

)٤٤(
احتیاج مارد باو بدهیم و ان عبارت است از ازادي حرکت و
سراهدهاي شخصي و تجربه ها و معلومات و میدان وسیع براي
انکشاف قوؤ فعالیت زیرا بچه میخواهد همه ٣بآ را بفهمد وکردار
ورفتار یادگیرد یعني بزرگ شود و کسال یابد ٠ ازینجا میتوان عي برد
باین که اگر ما بکوئیع زندگاني او را طوري مرتب سا٩به
که ابي احباجهاي او لا رفع کند ، او را ببهترین وجهي تربیت
کرد٠ایم و اوهم معلومات لازمه را یاد خواهد گرفت٠ آنوقت او
فوت و جمال و صحت بدني !یبا کسب خواهد کرد چرنکه بدن
هاد، انمان هم تنها از نان غذا و قوت نمیگیرد بلکه با خشنودي
مل نیر تغذیه میشود.

تصویر(۱۱): تصویر بازشناسی شده با ویژگی db1

)٤٤(
احتیاج مارد باو بدهیم و !ن عبارت است از !زادي حرکت و
سرامدهاي شخصي و تجربه ها و معلومات و میدان وسیع بر:ي
انکشاف قوؤ فعالیت زیرا بچه میخواهد همه چیز را بفهمد وکرمار
ورفتار یاد گیرد یعني بزرگ شود و کسال یابد ٠ ازینجا میتوان مي برد
باین فکند که اگر ما بگوشیم زندگاني او را طوري مرتب سا٩جع
که این احتیاجهاي او را رفع کند ، :و را ببهترین وجهي تربیت
کرد٠ایم و اوهم معلومات لازمه را یاد خواهدگرفت٠ آنوقت او
قوت و جمال و صحت بدني لین کسب خواهد کرد چونکه بدن
مادی، انسان هم تنها اؤ نان غذا و قوت نمیگیرد بلکه با خشنودي
مل نیز تغذیه میشود.

تصویر(۱۲): تصویر بازشناسی شده با ویژگی wdb1

به دلیل اینکه بازشناسی بر مبنای شکل کلی زیرکلمات بوده است، ابتدا در تصویر اولیه تمام زیرکلمات را شمارش نموده و سپس از روی تصویر اصلی در متن بازشناسی شده زیرکلمات اشتباه را پیدا کرده، شمارش نموده و از کل زیرکلمات کسر می نماییم. فرمول زیر نحوه‌ی انجام کار را نشان می‌دهد.

$$\text{درصد} = \frac{\text{تعداد زیرکلمات اشتباه} - \text{کل زیرکلمات}}{\text{کل زیرکلمات}} \times ۱۰۰$$

درصدهای بازشناسی بدست آمده در جداول زیر نمایش داده شده است . جدول (۵) جدول ارزیابی موجک سطح ۳ و جدول(۶) جدول ارزیابی بیته تبدیل موجک سطح ۲ را نمایش می‌دهد.

جدول (۵) جدول ارزیابی موجک سطح ۳

ویژگی‌ها	Haar	Db1	Db2	Coif1	Sym8
٠٠٠٨.*png*	٧٨٪	٨٠,٢٢٪	٧٩٪	٧٨٪	٦٧٪
٠٠١٧.*png*	٨٢,٦٪	٨٤,٥٢٪	٨٢,٧٨٪	٨٢,٦١٪	٧٦٪
٠٠٤٦.*png*	٨٤,٦٩٪	٨٥,٠٢٪	٨٤,١٩٪	٨٤,٨٥٪	٧٩,٣٪

جدول(۶) جدول ارزیابی بیته تبدیل موجک سطح ۲

ویژگی‌ها	WPHaar	WPDb1	WPDb2	WPCoif1	WPSym8
٠٠٠٨.*png*	٨٣,١٢٪	٨٥,٢٪	٨٤,٨٤٪	٨٣٪	٧٤,٦٧٪
٠٠١٧.*png*	٨٦,٩٥٪	٨٧,١٣٪	٨٦,٦٠٪	٨٤,٣٤٪	٧٦,٢٣٪
٠٠٤٦.*png*	٨٧,٣٪	٨٨,٦٨٪	٨٨,٣٥٪	٨٨٪	٨٠٪

۶- نتیجه‌گیری:

در این مقاله روش جدیدی با دقت و نرخ بازشناسی مناسب ٪۸۸ برای بازشناسی اسناد قدیمی چاپی فارسی ارائه شد. این روش

مبتنی بر ویژگی‌های تبدیل موجک و بسته تبدیل موجک در روش منهتن بوده است. نتایج مناسب الگوریتم نشان می‌دهد که الگوریتم ارائه شده به دلیل خاص و قدیمی بودن نوع فونت بکاررفته و همچنین استفاده از شکل کلی کلمات می‌تواند کارایی مناسبی در بازشناسی از خود داشته باشد. تفاوت موجود در این روش نسبت به روش های دیگر بدلیل الگوریتم ارایه شده و تعیین آستانه مناسب بوده است. کل کلمات مشابه و غیر مشابه بدست آمده ۳۷۸۶۱ بوده است، که توانستیم تعداد ۱۶۶۹ زیر کلمه غیر مشابه را جدا نموده و به عنوان واژه‌نامه استفاده نماییم.

مراجع

[۱] ح. نظام آبادی پور, ا. ا. کبیر, و ر. عزمی, "الگوریتم اصلاح شده جداسازی حروف در متون چاپی با برچسب زدن به کانتور بالایی کلمات," استقلال, شماره۱, ۱۳۸۳.

[۲] A. Al-Shoshan, "Arabic OCR based on image invariants." pp. ۱۵۰-۱۵٤.

[۳] H. Y. Abdelazim, and M. Hashish, "Arabic reading machine." pp. ۷۳۳-۷٤٤.

[٤] R. A. E. Kabir, "A new segmentation technique for omnifont Farsi text," *Pattern Recognition Letters*, vol. ۲۲, pp. ۹۷-۱۰٤, ۲۰۰۱.

[٥] ا. ابراهیمی, "معرفی، مفاهیم، تعاریف و اصطلاحات بازشناسی متن (OCR) " پژوهشنامه نویسه¬ خوان نوری (*OCR*) فارسی, چاپ اول, ص ۷۷-۱۰۹, ۱۳۸۸

[٦] S. Torabzadeh, and R. Safabaksh, "AUT-PFT: A real world printed Farsi text image dataset." pp. ۲٦۷-۲۷۲.

[۷] ا. ا. کبیر, و ا. ابراهیمی, "استفاده از شکل کلی زیر-کلمات چاپی در بازیابی تصویر مستندات و بازشناسی متون فارسی," رساله دکتری مهندسی برق-الکترونیک، دانشگاه تربیت مدرس, ۱۳۸٤.

[۸] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. ۹, no. ۲-٤, pp. ۱۲۳-۱۳۸, ۲۰۰۷.

[۹] A. Broumandnia, J. Shanbehzadeh, and M. R. Varnoosfaderani, "Persian/arabic handwritten word recognition using M-band packet wavelet transform," *Image and Vision Computing*, vol. ۲٦, no. ٦, pp. ۸۲۹-۸٤۲, ۲۰۰۸.

[۱۰] K. Huang, and S. Aviyente, "Mutual Information Based Subbband Selection for Wavelet Packet Based Image Classification." pp. ۲٤۱-۲٤٤.

[۱۱] A. A. Aburas, and Salem MA Rehiel, "Off-line omni-style handwriting Arabic character recognition system based on wavelet compression," *Arab Research Institute in Sciences & Engineering* vol. ۳, pp. ۱۲۳-۱۳٥, ۲۰۰۷.

[۱۲] م. شمسی, ع. ر. کناری, و س. شادروان, "روشی نو در تشخیص حروف," فصلنامه علمی – پژوهشی مهندسی برق مجلسی, شماره سوم, ۱۳۸۸.

[۱۳] H. Khosravi, and E. Kabir, "A blackboard approach towards integrated Farsi OCR system," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. ۱۲, no. ۱, pp. ۲۱-۳۲, ۲۰۰۹.

[۱٤] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:* ۱۱۰۰, ۱٤۰٤, ۲۰۱٤.

[۱٥] J. S. Walker, *A primer on wavelets and their scientific applications*: CRC press, ۲۰۰۸.