Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ingeniería en Ciencias y Sistemas

LA APLICACIÓN DEL RECONOCIMIENTO ÓPTICO DE CARACTERES EN LA INDUSTRIA DEL PROCESAMIENTO MASIVO DE INFORMACIÓN

EDDY ROLANDO VELÁSQUEZ CASTILLO

ASESORADO POR EL ING. CRESENCIO GERTRUDIS CHAN CANEK

Guatemala, Marzo de 2005

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

LA APLICACIÓN DEL RECONOCIMIENTO ÓPTICO DE CARACTERES EN LA INDUSTRIA DEL PROCESAMIENTO MASIVO DE INFORMACIÓN

TRABAJO DE GRADUACIÓN

PRESENTADO A JUNTA DIRECTIVA DE LA

FACULTAD DE INGENIERÍA

POR

EDDY ROLANDO VELÁSQUEZ CASTILLO

ASESORADO POR EL ING. CRESENCIO GERTRUDIS CHAN CANEK

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, MARZO DE 2006

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	
VOCAL II	Lic. Amahán Sánchez Álvarez
VOCAL III	Ing. Julio David Galicia Celada
VOCAL IV	Br. Kenneth Issur Estrada Ruiz
VOCAL V	Br. Elisa Yazminda Vides Leiva
SECRETARIA	Inga. Marcia Ivonne Véliz Vargas

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	**Ing. Sydney Alexander Samuels Milson**
EXAMINADOR	Ing. Luis Alberto Vettorazzi España
EXAMINADOR	Inga. Virginia Tala
EXAMINADOR	Ing. Edgar Santos
SECRETARIO	Ing. Carlos Humberto Pérez Rodríguez

HONORABLE TRIBUNAL EXAMINADOR

Cumpliendo con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

LA APLICACIÓN DEL RECONOCIMIENTO ÓPTICO DE CARACTERES EN LA INDUSTRIA DEL PROCESAMIENTO MASIVO DE INFORMACIÓN

Tema que me fuera asignado por la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería en febrero de 2004.

Eddy Rolando Velásquez Castillo

Guatemala 18 de Mayo de 2005

Ing. Carlos Alfredo Azurdia Morales
Coordinador de Privados y Trabajos de Graduación
Escuela de Ingeniería en Ciencias y Sistemas
Facultad de Ingeniería
Universidad de San Carlos de Guatemala

Ing. Azurdia

Por medio de la presente hago de su conocimiento que he tenido a bien revisar el trabajo de graduación de EDDY ROLANDO VELASQUEZ CASTILLO, titulado ”LA APLICACIÓN DEL RECONOCIMIENTO ÓPTICO DE CARACTERES EN LA INDUSTRIA DEL PROCESAMIENTO MASIVO DE INFORMACIÓN”, por lo cual me permito recomendar dicho trabajo final para la respectiva revisión por parte de la comisión de trabajos de graduación de la escuela de Ciencias y Sistemas.

Sin otro particular, me suscribo atentamente,

Ing. Cresencio Gertrudis Chan Canek
ASESOR

ÍNDICE

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

GLOSARIO

Algoritmo	Descomposición en pasos u operaciones fundamentales de cualquier operación de cálculo o proceso analógico para su resolución optima.
Bottom-Up	De abajo hacia arriba, metodología que va de lo mas específico a lo más general.
Batch	Colección de cosas o personas que serán manejadas juntas.
Escáner	Aparato que se utiliza para la exploración con el fin de obtener diferentes imágenes de una región.
ICR	*Inteligent Character Recognition;* en español; Reconocimiento Inteligente de Caracteres.
Lexicon	Conocimiento de palabras; libro de referencia conteniendo un listado alfabético de palabras con información acerca de ellas; diccionario.

OCR	*Optical Character Recognition;* en español: Reconocimiento óptico de caracteres.
Píxel	El más discreto componente de una imagen o figura en un monitor de puntos a colores.
Segmentación	Dividir en segmentos o pedazos.
Semántica	Referente al significado de las palabras.
Script	Una escritura o sistema en particular.
Tabla de Hash	Estructura de datos que permite la búsqueda y almacenamiento de información.
Token	Una instancia individual de un tipo de símbolo.
Top-down	De arriba abajo; metodología que va de lo general a lo específico.

OBJETIVOS

General:

1. Enriquecer el conocimiento general de cómo funciona el reconocimiento óptico de caracteres, para formar un punto de vista que permita al lector conocer cuáles factores debe considerar al utilizar esta tecnología.

Específicos

1. Dar a conocer los principios básicos bajo los cuales funciona el OCR.
2. Mostrar el rol de la inteligencia artificial y del procesamiento del lenguaje natural, así como las limitantes que se presentan al aplicarlo al OCR.
3. Proporcionar información acerca del rendimiento de este tipo de aplicaciones en escenarios reales que permitan tomar decisiones al elegir este tipo de tecnología.
4. Conocer cuáles son las técnicas tradicionales de procesamiento de datos y cuáles son los factores que influyen en el uso de OCR en lugar de dichas técnicas.
5. Dar a conocer las experiencias de las aplicaciones del reconocimiento óptico de caracteres en Guatemala.

INTRODUCCIÓN

En un mundo en que la población aumenta día a día y de la cual tarde o temprano necesitaremos conocer sus características y gustos, procesar sus impuestos, o sus envíos de correo, su historial médico, quizás conocer sus habilidades técnicas, imagine por un momento el tiempo que le lleva a alguien procesar este tipo de información. Ahora bien, piense, ya no en cientos, sino en miles o quizás millones de documentos, de los cuales se desea obtener información, los costos de procesar esta información de una forma tradicional (digitación y revisión) se harían casi prohibitivos, eso sin mencionar el tiempo que llevaría completar dicho procesamiento. Es ahí donde el reconocimiento óptico de caracteres juega un papel importante, al proveer de herramientas que permitan el procesar información en poco tiempo y de forma menos costosa que el procesamiento de datos tradicional. Pero es necesario tomar en cuenta que habrán factores externos que podrán afectar el rendimiento de un sistema de este tipo, por lo cual se hará necesario conocer cómo prevenir y corregir estos errores.

En la fase de corrección de errores es en donde la Inteligencia Artificial jugará un papel por demás importante, al permitirnos mediante el procesamiento de lenguaje natural identificar el sentido semántico de una frase,

ampliando el rango de procesamiento más allá de las marcas, si bien en esta fase también se encontrarán problemas, será nuevamente la inteligencia artificial la que nos proveerá de una solución que facilitará reconocer imágenes que representen algún tipo de letras.

Se pretende con todo esto, mostrar las bases bajo las cuales funciona el OCR comercial, y a la vez mostrar casos prácticos en los cuales se podrá evaluar el desempeño de esta tecnología en ambientes de trabajo real; y con ello mostrar que el uso de OCR puede llegar a ser algún día como en algunos otros países, una forma bastante factible de procesar datos.

1. RECONOCIMIENTO ÓPTICO DE CARACTERES

1.1 Introducción

El Reconocimiento Óptico de Caracteres (OCR *Optical Character Recognition*, por sus siglas en inglés) es una técnica de gran interés desde que Jacob Rabinow iniciara a investigar este campo en finales de 1940 [1].

En sus inicios, las máquinas que utilizaban OCR eran máquinas mecánicas con altas tasas de error al interpretar y reconocer caracteres, con el pasar de los años la necesidad de procesar información impresa o escrita hacía que dichas máquinas se viesen obsoletas y no se prestasen para realizar las tareas de procesamiento de información, esto dio paso a una nueva generación de dispositivos OCR que emplean distintos algoritmos, diseñados para trabajar con computadoras. Aunque ninguno de estos algoritmos es confiable en un 100 %, los mas populares son bastante eficaces y rápidos, pero aun cometen errores que para una persona pueden parecer inaceptables.

Estos errores, se deben en la mayoría de los casos a los algoritmos que se utilizan, que si bien desde el punto de vista computacional son bastante exactos, se basan, mas que en interpretar trazos o líneas y sus relaciones, como lo haría una persona, en medir mediante avanzadas funciones matemáticas, que van desde Proyecciones Transformadas en Anillo (TRP) hasta la Transformada de Fourier de las Proyecciones Verticales y Horizontales de un carácter.

Puede que algunos errores, en cambio parezcan aceptables, pero bajo cualquier punto de vista hemos de recordar que el objetivo principal del uso del OCR es facilitar la comunicación con los humanos.

1.2 Segmentación, Vectorizacion y Reconocimiento

1.2.1 Segmentación de imágenes

¿En qué consiste el Reconocimiento Óptico de Caracteres? La principal tarea del reconocimiento óptico de caracteres consiste en identificar de entre un grupo de caracteres, el carácter que un arreglo de píxeles interpreta, este arreglo de píxeles tiene como origen una imagen que pudo ser escaneada u obtenida por algún otro medio físico. De acuerdo con el algoritmo que se utilice, el arreglo de píxeles puede variar según el tipo de caracteres que se interpreta, así como la calidad de la imagen origen.

La mayoría de dispositivos OCR utiliza algoritmos en los cuales se evalúa un punto de referencia, este punto de referencia consiste en un punto de principal atención para el ojo de quien lo contempla, puede ser una esquina, la intersección de varias líneas o el centro de alguna imagen.

En este caso el punto de referencia, puede ser cualquiera de las uniones de un carácter, estas uniones son quienes le dan sentido a este carácter, por ejemplo, la Z es un grupo de líneas, que no tendrían sentido, si no fuese por los puntos por los cuales está unida.

Como parte del procesamiento óptico tenemos la segmentación, que consiste en separar la imagen en porciones o cuadros que la representan, los algoritmos se encargan de detectar que casillas o segmentos alrededor de un punto de referencia están o no ocupadas.

Debido a que los caracteres occidentales pueden ser representados en arreglos de 8 x 8, es que después de algunas reducciones de los datos obtenidos, se pueden definir hasta 256 distintas combinaciones, cualquiera de estas 256 combinaciones será comparada contra las 256 combinaciones posibles píxeles vecinos. (ver Tabla 1)

El proceso de comparar punto por punto, es llamado Agrupamiento de cadena simple (*Single Link Clustering*) que utiliza la distancia mínima entre 2 grupos dados y todos los puntos contiguos[1], donde un grupo o *cluster* es definido como una serie de puntos representados por la siguiente formula:

$$d_{\min}\left(D_i, D_j\right) = \min_{\substack{x \in D_i \\ x \in Dj}} \left\| x - x' \right\|$$

[1] Barney Smith, Elisa H. Document Scanning Defect Analysis using Bilevel Image Features, Electrical Engineering Department, Boise State University. EEUU.

1.2.2 Vectorización como método de procesamiento de imágenes para el reconocimiento de caracteres

Los algoritmos utilizados para reducir la información obtenida de arreglos de mayores dimensiones son conocidos como algoritmos de adelgazamiento, estos están basados en la VECTORIZACION la cual es usada como un método de pre -procesamiento para el reconocimiento óptico de caracteres. Las técnicas de vectorización han sido desarrolladas en varios dominios y numerosos métodos han sido propuestos e implementados.

Los métodos de adelgazamiento usualmente utilizan procesos iterativos de erosión de fronteras para remover los píxeles más externos hasta que queda un solo píxel base o esqueleto; un proceso de seguimiento de líneas es utilizado para enlazar el esqueleto a una cadena de píxeles.

Un procedimiento de aproximación poligonal es aplicado para convertir la cadena de píxeles a un vector, que puede ser desde una barra simple o un polígono (una cadena de 2 o más barras), la principal ventaja de este método es el mantener la conectividad de las líneas, las desventajas consisten en perdida del ancho de la información, distorsión en las uniones e ineficacia de tiempo, pero con un buen algoritmo estas deficiencias pueden reducirse considerablemente.

Tabla I: Enumeración de posibles píxeles vecinos para un arreglo de 8 x 8

															
0	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240
															
1	17	33	49	65	81	97	113	129	145	161	177	193	209	225	241
															
2	18	34	50	66	82	98	114	130	146	162	178	194	210	226	242
3	19	35	51	67	83	99	115	131	147	163	179	195	211	227	243
4	20	36	52	68	84	100	116	132	148	164	180	196	212	228	244
5	21	37	53	69	85	101	117	133	149	165	181	197	213	229	245
6	22	38	54	70	86	102	118	134	150	166	182	198	214	230	246
7	23	39	55	71	87	103	119	135	151	167	183	199	215	231	247
8	24	40	56	72	88	104	120	136	152	168	184	200	216	232	248
9	25	41	57	73	89	105	121	137	153	169	185	201	217	233	249
10	26	42	58	74	90	106	122	138	154	170	186	202	218	234	250
11	27	43	59	75	91	107	123	139	155	171	187	203	219	235	251
12	28	44	60	76	92	108	124	140	156	172	188	204	220	236	252
13	29	45	61	77	93	109	125	141	157	173	189	205	221	237	253
14	30	46	62	78	94	110	126	142	158	174	190	206	222	238	254
15	31	47	63	79	95	111	127	143	159	175	191	207	223	239	255

Como resultado de este pre - procesamiento mediante algoritmos de adelgazamiento se obtienen resultados como los presentados en la Figura 2.a, en los que se puede observar que la eliminación de los puntos más externos de la imagen han hecho que esta se vea menos definida que la inferior, pero debido a que presenta los primeros rasgos de pixelizado, con varias pasadas por el mismo algoritmo podrá verse definida en una de las combinaciones de la Tabla I.

Figura 1. Puntos de una imagen antes de aplicarle un algoritmo de adelgazamiento

Figura 2. Texto al que se le ha aplicado un algoritmo de adelgazamiento

1.2.3 Algoritmos de Segmentación de paginas

Los algoritmos de segmentación de páginas pueden ser catalogados en tres clases: los de metodología *top-down,* los *bottom-up* y los híbridos, a continuación se describen 3 de los algoritmos mas difundidos y que representan cada una de las metodologías antes descritas, el algoritmo de corte X-Y (metodología *top - down*), el algoritmo de Docstrum y el algoritmo de Voronoi (metodología *bottom-up*).

.

1.2.3.1 Algoritmo de segmentación de pagina en corte X-Y

Este algoritmo se basa en árboles y algoritmos del tipo *top-down*, la raíz de este árbol representa la imagen de la pagina inicial del documento, y todas las hojas nodo representan la segmentación final. Aunque este documento es fácil de implementar, únicamente puede trabajar con documentos que no presenten desnivel entre cada una de las líneas del documento y con zonas rectangulares.

Los pasos de este algoritmo son:

1. Crear las tablas horizontales y verticales de prefijos H_X y H_Y como se describe a continuación:

$$H_X[i][j] = \#\{p \in D(I) \mid X(p) = j, Y(p) \le i, I(p) = 1\},$$
$$H_Y[i][j] = \#\{p \in D(I) \mid X(p) \le j, Y(p) = i, I(p) = 1\},$$

Donde $D(i) \subseteq Z^2$ representa al dominio de la imagen *I* , *I(p)* es el valor binario de la imagen en el píxel *p*, *X(p)* e *Y(p)* son las coordenadas X – Y del píxel *p* respectivamente.

2. Inicializar el árbol con la imagen completa del documento como nodo raíz. Para cada nodo haga lo siguiente
   a. Calcule el histograma del perfil de la proyección del píxel en X – Y para el nodo actual de la siguiente forma:

$$HIS_X[i] \leftarrow H_X[Y_2(Z)][i] - H_X[Y_1(Z)][i],$$
$$HIS_X[j] \leftarrow H_Y[j][X_2(Z)] - H_Y[j][X_1(Z)],$$

   Donde *Z* es la zona correspondiente al nodo actual, y los puntos de la superior izquierdo e inferior derecho de la zona están representados por $(X_1(Z), Y_1(Z))$, $(X_2(Z), Y_2(Z))$ respectivamente

   b. Reduzca la zona actual hasta que esta encierre de forma ajustada el cuerpo de la zona, es entonces que los umbrales para remover ruido T_X^N y T_Y^N son utilizados para clasificar y remover el ruido de los píxeles de fondo, como se asume que dichos píxeles están distribuidos de forma uniforme, los valores para remover el ruido del fondo para un nodo especifico son medidos de forma lineal para el ancho y alto de la zona de trabajo actual

   c. Repita el paso 2ª

   d. Obtenga los valles V_X y V_Y que presenten mayor amplitud en los histogramas HIS_X y HIS_Y de proyección de perfil para X,Y

e. Si $V_X > T_X^C$ o $V_Y > T_Y^C$,donde T_X^C y T_Y^C representan la amplitud de dos umbrales, divida en el punto medio de los mas amplios V_X y V_Y, genere dos nuevos nodos hijos, de lo contrario haga del nodo actual una hoja.

1.2.3.2 El algoritmo de segmentación de paginas de Docstrum

El algoritmo de Docstrum es un algoritmo de segmentación de paginas del tipo *bottom-up* que puede trabajar en documentos que presentan diseños no estándares y ángulos retorcidos. Este algoritmo no esta diseñado para manejar regiones que no sean de texto y regiones que presenten tipos de letra irregular o espacios que las fragmenten. Peor aun, no se desempeña bien cuando las imágenes del documento contienen caracteres dispersos.

1. Obtenga todos los componentes conectados (C_i^8) con algún algoritmo eficiente de dos pasadas

2. Elimine todo el ruido y componentes conectados no texto, utilizando los umbrales de bajo (*l*) y alto (*h)*

3. Separe los componentes C_i^8 en dos grupos, uno con los caracteres dominantes y otro con los títulos, encabezados, etc. Un parámetro f_4 controla como estos se compactan.

4. Encuentre los *K* vecinos mas cercanos, $NN_K(i)$ para C_i utilizando una aproximación ordenada

5. Calcule la distancia y el ángulo para cada C_i y sus K vecinos cercanos (ρ^i_j , θ^i_j), tal que $j \in NN_K(i)$.

6. Calcule un histograma de la distancia del vecino cercano alineado del siguiente conjunto $W\rho$: $W\rho = \{\rho^j_j \mid j \in NN_K(i), y \ -\theta_h \leq \theta^i_j \leq \theta_h\}$ donde θ_h es el ángulo horizontal del umbral de tolerancia. Estime el espaciamiento entre caracteres de una misma línea *cs* así como la localización del punto mas alto del histograma.

7. Calcule un histograma con la distancia del vecino mas cercano entre líneas del conjunto $B\rho$: $B\rho = \{\rho^j_j \mid j \in NN_K(i), y\ 90^o - \theta_\upsilon \leq \theta^i_j \leq 90^o + \theta_\upsilon\}$ donde θ_υ es el ángulo vertical del umbral de tolerancia. Estime el espaciamiento entre líneas *ls* así como la localización del punto mas alto del histograma.

8. Realice un acercamiento transitorio de los vecinos más cercanos de la misma línea, emparejándolos a fin de obtener líneas de texto L_{iS} utilizando los umbrales de distancia de los vecinos más cercanos de la línea, $T_{cs}=f_i * cs$

9. Realice un acercamiento transitorio sobre L_{iS} para obtener bloques estructurales o zonas Z_{iS} utilizando un umbral paralelo de distancia $T_{pa}=f_{pa} * cs$ y el umbral perpendicular de distancia $T_{pe}=f_{pe} * ls$ las distancias perpendiculares y paralelas son calculadas como distancias borde – borde, no centro – centro.

1.2.3.3 Algoritmo de segmentación de páginas basado en el diagrama de Voronoi

Este algoritmo, también de los de tipo *bottom-up* esta basado en el diagrama de Voronoi, este método puede trabajar sobre imágenes de páginas de documentos que tienen diseños no estándar, ángulos retorcidos de forma arbitraria, y líneas de texto no lineares. Un conjunto de segmentos conectados por líneas son utilizados como áreas de unión de texto, pero presenta algunos de los problemas que presenta el algoritmo de Docstrum. A continuación se describen sus pasos:

1. Marque los componentes conectados. Tome puntos de muestra en sus bordes. El parámetro *sr* controla el número de puntos de muestra que serán utilizados.
2. Elimine el ruido de los componentes conectados utilizando el máximo umbral de zona de ruido *um*, el máximo umbral de ancho C_w, el máximo umbral de altura C_h, el máximo umbral de radio C_r para todos aquellos componentes conectados.
3. Haga un diagrama de Voronoi para cada uno de los componentes conectados, utilizando los puntos de muestra de sus bordes.
4. Elimine los bordes superfluos del diagrama de Voronoi, para generar las fronteras de las zonas de texto, de acuerdo al criterio de área – radio.
5. Elimine las zonas de ruido utilizando el umbral mínimo de área A_z para todas las áreas, y utilizando el umbral mínimo de área A_l y el vertical de radio B_r para las zonas verticales y prolongadas.

1.3 Relacionando letra por letra, palabra por palabra.

El término de asociación de palabras es utilizado en el sentido particular de la literatura psico – lingüística. Este concepto implica que ciertas palabras pueden ser relacionadas con otras de acuerdo al contexto que abarquen, es una práctica común en la lingüística el clasificar palabras no solo en base de su significado, sino en base a sus coincidencias con otras palabras, la búsqueda de palabras coincidentes no es fácil, pero con la ayuda de sistemas computacionales estas tareas se vuelven mas manejables y fáciles.

En las secciones anteriores se ha pasado por las distintas fases que tiene el reconocimiento óptico de caracteres, desde el pre procesamiento que toma la imagen y la divide en secciones pequeñas o fase de segmentación, de ahí a donde las sub imágenes de los caracteres se dividen y se analiza la relación de ellas, luego pasa a la fase de reconocimiento en donde se reconoce a toda la imagen como un carácter de acuerdo a una base de datos en la que se almacenan distintas imágenes para cada carácter, y ahora en la fase de post procesamiento en donde se relaciona carácter con carácter para formar palabras y estas para formar oraciones de acuerdo al tipo de aplicación que se tenga, y el lenguaje que esta utilice.

1.3.1 El lenguaje como modelo

Entre los diferentes niveles en los cuales puede ser modelado el lenguaje, el más bajo es el que esta a nivel de palabras, incluyendo restricciones léxicas en las relaciones secuenciales de los caracteres en cada palabra. El siguiente nivel es el de oraciones, el cual toma en cuenta restricciones sintácticas y semánticas en la secuencia de palabras o categorías de palabras dentro de una oración (o un campo, dentro de un formulario, en una aplicación para el procesamiento de formularios).

Los niveles mas altos consideran la amplitud del contexto y requieren un conocimiento específico para el dominio de la aplicación. A los modelos a nivel de palabra y oración comúnmente se les aplican métodos de búsqueda en diccionarios, técnicas basadas en distancias de edición, Cadenas de Markov y otros modelos de transición, una meta común para el OCR es garantizar que las palabras u oraciones producidas por OCR son correctas en el sentido que ellas pertenezcan al lenguaje utilizado, dando lugar a dos puntos de vista para la resolución y búsqueda de palabras, que cumplan con las restricciones de su lenguaje.

El punto de vista determinista es a menudo aplicado a lenguajes sin mayores complicaciones y requiere en su forma mas simple la compilación de un diccionario (una lista finita de *tokens* que pueden ser palabras, oraciones, etc.)

En este caso, un *token* resultado de un algoritmo OCR y al que se le conocerá como cadena reconocida, se dirá que pertenece al lenguaje con un cierto porcentaje de similitud al *token* mas parecido en el diccionario, el *token* será devuelto como la cadena corregida en respuesta al *token* de entrada; cuando no es posible construir una lista con todas las posibles cadenas válidas, una representación no enumerada puede ser utilizada como una gramática o conjunto de prefijos, raíces de palabras, sufijos, etc.

En un modelo no determinista la cadena corregida no es forzada a pertenecer necesariamente a una lista explícita o implícita de *tokens* enumerados, en cambio la cadena reconocida es analizada y una cadena corregida es construida mediante maximizaciones de la probabilidad que cada símbolo o sub secuencia de símbolos pertenezca al lenguaje. En otras palabras las cadenas corregidas son una versión de las cadenas reconocidas que conforman de manera mas cercana las restricciones del modelo del lenguaje, independientemente del tipo de modelo y lenguaje que utilice, este tendrá una serie de pasos que son los mas comunes y utilizados.

1.3.2 Búsqueda y Asociación

Empezando por el Análisis de Contexto el cual es hecho por aplicaciones iterativas de varios módulos simples los cuales intentan asignar etiquetas a los segmentos mediante distintas reglas.

Cada asignación tentativa es evaluada usando un radio vector posición el cual es el número de palabras validas de un *lexicon* o diccionario con varios patrones de palabras que contiene. Una patrón de palabra es la raíz de una palabra, el cual puede contener una mezcla de elementos etiquetados y no etiquetados.

Solamente uno se tomará forzándolo a coincidir, incluso si hay múltiples coincidencias para el mismo patrón, la interpretación de palabras son construidas progresivamente a partir de asignaciones previamente aceptadas, los módulos más utilizados para esta comparación - asignación son[2]:

- Asignación conjunta: se toman los 3 segmentos mas largo y se trata de asignarlos a cada trío o las ocho letras mas comunes (según el idioma), de donde se selecciona el trío que maximice el número de el número de coincidencias dentro del diccionario entre aquellas que tengan al menos dos de las ocurrencias de estos tres segmentos.

- Coincidencia única: a continuación cada patrón de palabra conteniendo al menos uno de los segmentos no etiquetados es revisado, asignando una o varias letras al segmento vacío, para ver si alguna de las asignaciones coincide con alguna única entrada del diccionario, esto se hace de forma iterativa hasta que no se encuentra ninguna otra coincidencia.

[2] Brown, Eric W. Character Recognition by Feature Point Extraction EEUU.

- Coincidencia más encontrada: si aun hay segmentos sin etiquetar, entonces el algoritmo asigna cada letra en turno a cada uno de los segmentos no asignados y revisa si las asignaciones resultan en el mas alto radio. Si el mejor radio es al menos de 0.75 y el segundo mejor no es muy cercano (al menos 0.1 menor) entonces su asignación es confirmada.

- Asignación Verificada: finalmente cada asignación es verificada intentando reemplazarla con cada letra del alfabeto, si el radio mejora y mas de una palabra contiene al menos dos letras, entonces la etiqueta es reemplazada, y la verificación continua. Se aplican algunas precauciones, como la de evitar que segmentos que aparecen solo una vez como una palabra de una única letra reciban otras, a menos que el contexto de las otras palabras lo justifique.

Las búsquedas descritas anteriormente van desde métodos que incluyen estructuras de datos que realizan búsquedas de manera eficaz dentro de un diccionario.

La forma mas simple es una lista de palabras ordenadas de forma alfabética, desafortunadamente se requiere técnicas específicas e índices para hacer búsquedas en ellas. Lo mismo puede decirse de otras estructuras de datos genéricas como tablas de Hash, árboles de búsqueda, etc. Otras técnicas mas especificas tales como métodos que tratan de beneficiar el alto rendimiento de los algoritmos de búsqueda avanzada, generando un número de cadenas vecinas y buscando entre ellas la que más se ajuste.

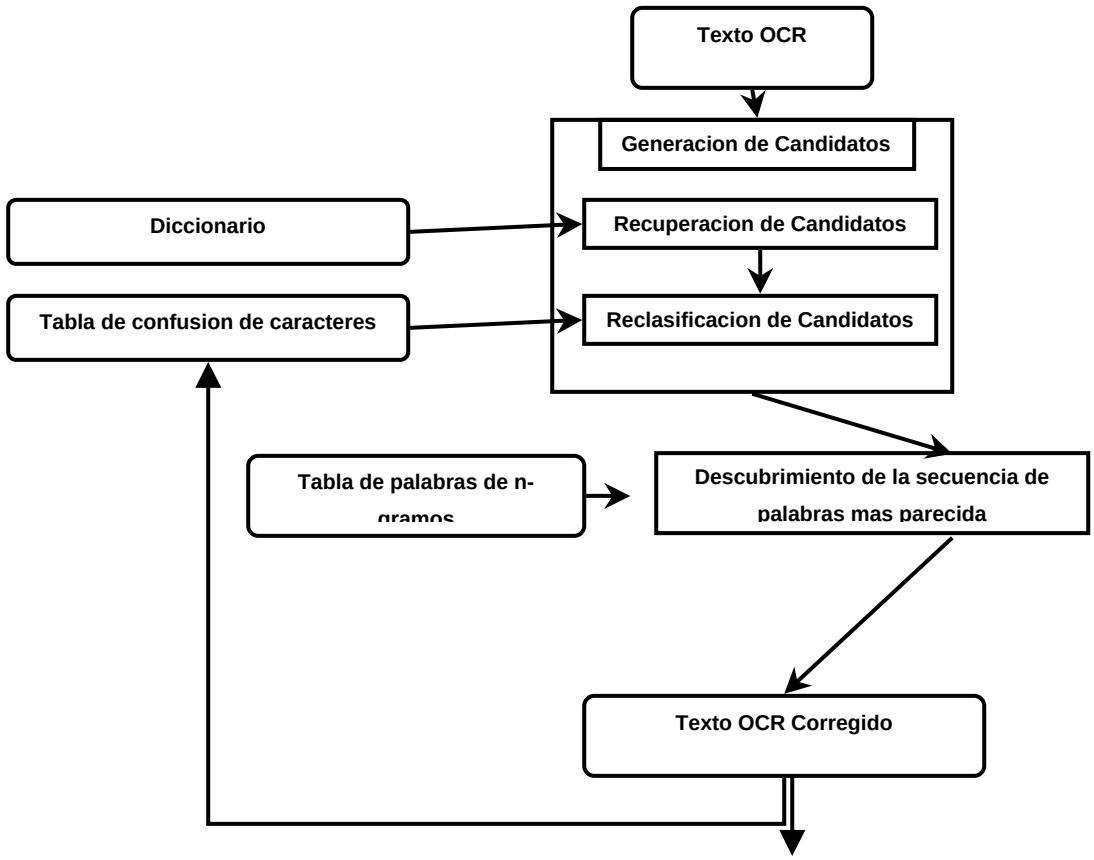
Lamentablemente el número de vecinos puede crecer exponencialmente con el número de símbolos de la cadena, además de las restricciones heurísticas y probabilísticas que se han impuesto. En otros casos el diccionario es dividido de acuerdo a diferentes criterios, como el largo de las palabras, el primer símbolo, etc., reduciendo el tiempo de búsqueda por un factor constante.

Otro grupo interesante de técnicas esta basada en técnicas de búsqueda rápida, considerando las cadenas como puntos con diferencias de espacio y realizando técnicas de reducción de dimensiones, búsquedas de árboles, o búsquedas rápidas basadas en las propiedades de las medidas, todas estas técnicas tienen como objetivo relacionar caracteres en palabras, y de acuerdo al modelo estocástico o no estocástico, en relacionar palabras para formar e interpretar oraciones.

1.3.3 Pérdida de Información y su solución

Casi desde el momento en que las computadoras se hicieron de uso general apareció un concepto conocido como *paperless office* u “oficina sin papeles”, concepto que tenía como objetivo general librar al mundo de enormes cantidades de información escrita o impresa, a cambio de información de fácil acceso y almacenamiento, la conversión de estos documentos impresos a formatos electrónicos se haría mediante el uso de escáneres que producirían una imagen digital del documento.

Figura 3. Arquitectura de un sistema OCR



Fuente: Xiang Ton, *A Statistical Approach to Automatic OCR Error Correction in Context*

Aunque algunas veces esta imagen digital es lo único que se necesita a menudo se requiere que esta imagen pueda ser interpretada y procesada, en cualquiera de los casos la imagen producida por el escáner es diferente a la que aparece impresa, y al obtener texto de una imagen a través de OCR o de diseño asistido por computadora (CAD).

El texto que obtengamos puede no ser exactamente el que vemos impreso. La degradación que el texto presenta a partir de la imagen que le dio origen puede darse en cualquiera de las fases de vida del documento**,** En la figura 3, se presentan los pasos que sigue el procesamiento óptico de una imagen, en cada proceso puede aparecer un factor que resulte determinante para el resultado que se obtenga del OCR.

1.4 Errores al nivel de caracteres y sus primeras soluciones

Cuando las personas leen el documento producido, algunas imperfecciones son ignoradas, debido a la capacidad cerebral de recuperarse de errores, pero cuando la calidad del texto se ve seriamente afectado por el nivel de degradación de la imagen es cuando este requiere mayor atención. La mayoría de errores que una persona nota al leer texto obtenido mediante OCR se debe a que el sistema OCR fue incapaz de clasificar un carácter a partir de la imagen.

A partir de un estudio realizado por el *Information Science Research Institute* (ISRI)*,* en el cual se trataba de medir la calidad de varios sistemas OCR comerciales, se obtuvieron como resultados que para una página que contenía 2000 caracteres, el 99% de exactitud al obtener el texto representaba que aun existían por lo menos 20 errores por página, cuando la exactitud era solamente del 95% había alrededor de 100 errores por página, a partir de estos números se puede observar el porqué la pérdida de información al utilizar OCR sigue siendo de bastante interés.

Los errores como resultado del uso de OCR, son causados por diversos factores, para su evaluación, han sido agrupados en 4 categorías principales:

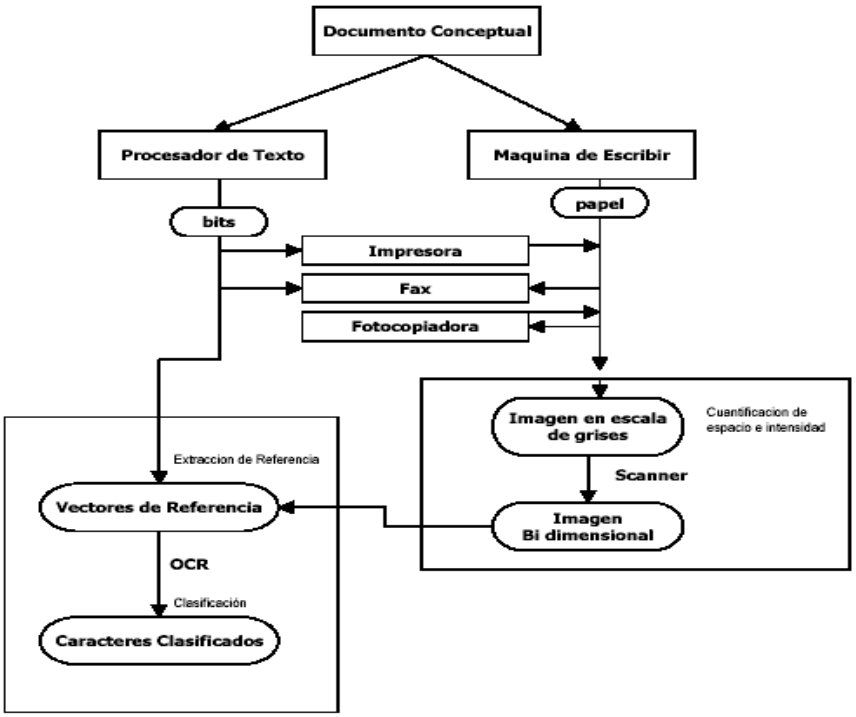
- Defectos de Imagen: estos defectos ocurren entre el proceso de impresión y el proceso del reconocimiento óptico de caracteres.

- Símbolos similares: todos los dispositivos OCR reconocen caracteres principalmente por su forma, pero debido a que existen caracteres con formas similares (por ejemplo 5 y S), esto puede causar confusión al interpretar la imagen.

- Puntuación: el uso de letras mayúsculas y puntuación son guías en material escrito, como el acento y las pausas son en el habla, pero en el reconocimiento óptico de caracteres, puede implicar confusión entre símbolos o bien interpretarse como símbolos desconocidos.

- Tipografía: resultados de la propia naturaleza humana en la que la variedad nos lleva a modificar la apariencia de texto escrito, al escáner o copiar con distintos tipos de fuente y otros efectos.

Figura 4. Flujo de un documento a través del uso de OCR2.



Fuente: Barney Smith, Elisa H., Document Scanning Defect Analysis using Bilevel Image Features

La habilidad de caracterizar las degradaciones que ocurren al convertir un documento en papel a una imagen digital a través del escaneado es un paso importante para mejorar la exactitud de este tipo de procesamiento, se han propuesto y desarrollado varios métodos que permiten calcular la cantidad o nivel de defectos en un documento, que en determinado momento permiten tomar la decisión de si es mejor ingresar el texto o bien utilizar un paquete OCR y luego hacer las correcciones necesarias.

Los errores en la conversión binaria de la imagen y el ancho PSF (*Point Spread Function* o Funcion de Puntos Dispersos) causan que los trazos que forman la imagen se reduzcan o adelgacen, y en otros casos que se aumente su grosor, dando como resultado imágenes incompletas o que se entrelazan, estas interfieren con la segmentación, siendo esta una de las mayores fuentes de error en los sistemas OCR, para mejorar su exactitud se han combinado modelos de procesos de degradación junto a estimados de los parámetros en estos modelos.

Uno de estos modelos conocido como el Modelo de Baird consiste en una convolusión del sensor de la función de punto disperso seguido (PSF) de un re - acondicionamiento para crear una imagen bidimensional, esto es hecho agregando un parámetro para muchos otros defectos como ruido, movimiento, desnivel entre el texto y la barra de escaneo.

Existen otros modelos que plantean soluciones parecidas a estas (Modelos de Kanungo, Sarkar, Loce y Wolberg) pero que aun no dan solución definitiva a los errores que presenta el utilizar un sistema OCR, siendo aún una opción el decidir si es preferible ingresar el texto en forma manual o utilizar un sistema OCR y realizar también de forma manual las correcciones al documento producido.

1.4.1 Corrección a nivel de palabras

Hasta el momento hemos podido definir los errores a nivel de caracteres, pero ¿Qué pasa con las palabras?, Existen métodos para la corrección del post procesamiento de OCR, los errores de este post procesamiento pueden ser debidos a la mala asignación de palabras comparadas o bien a que el diccionario no sea muy exacto.

El Algoritmo de Viterbi ha sido utilizado para la corrección de errores estocásticos con excelentes resultados en el reconocimiento de patrones, La extensión de la corrección de errores implica la adición de nuevos parámetros de acuerdo a las reglas de error. A nivel de palabras y oraciones son mas utilizados los modelos de error basados en operaciones de edición, inserción y eliminación. El costo de estas operaciones puede variar, siendo el mas bajo el de substitución, mas adelante se profundiza mas este tema.

Si bien los algoritmos que se utilizan en el reconocimiento óptico de caracteres, ni los empleados en relacionar caracteres para formar palabras son aun 100% exactos, se han descrito métodos que permiten definir una serie de pasos para corregir los errores resultado del post procesamiento:

- Leer una oración del texto OCR
- Obtener al menos M candidatos del diccionario por cada posible error, re clasifique los M candidatos por sus probabilidades condicionales de error. Mantenga únicamente a los N candidatos para el siguiente paso del procesamiento (N esta dado proporcionalmente a M).

- Utilice el algoritmo de Viterbi para lograr que la mejor secuencia de palabras para las cadenas en la oración.

Este método requerirá varias pasadas para corregir el texto OCR, en la primera pasa el sistema no tendrá información acerca de las probabilidades de confusión de caracteres, este asumirá que el carácter esta reconocido de forma correcta. En las siguientes pasadas el sistema distribuirá el resto de la probabilidad de forma uniforme entre los otros eventos, en cada paso que de se retro - alimentará, generando una tabla con las probabilidades para la confusión de caracteres, comparando el texto OCR con el texto OCR corregido en la última pasada (ver figura No. 3).

2. RECONOCIMIENTO ÓPTICO DE CARACTERES Y EL PROCESAMIENTO DEL LENGUAJE NATURAL

2.1 Introducción

El procesamiento del lenguaje natural nacido a finales de la década de 1950 y principios de 1960 busca representar objetos lingüísticos de forma que estos no sean representados únicamente como números sino como objetos lingüísticos con un significado especial, dando una nueva visión de las computadoras, ya no solo como máquinas que son buenas en procesos aritméticos, sino también como máquinas capaces de manipular símbolos e incluso objetos mas complejos tales como palabras, oraciones, árboles o redes.

Si bien el conocimiento humano puede ser representado en forma de procedimientos, es en realidad la esencia de un texto que al ser procesado por el cerebro humano, el que tendrá infinidad de formas para asociar, por ejemplo, una palabra con un lugar, o una imagen con una palabra pero la forma de hacerlo resulta ser mucho mas complicada que una serie de comandos de programación, y hace notorio el hecho de que el campo de la inteligencia artificial aun tiene mucho camino por recorrer.

2.2 ¿Es posible enseñar a un sistema como relacionar palabras?

2.2.1 Procesamiento del Lenguaje Natural y la gramática como parte de este procesamiento

Si vemos un programa que realice alguna tarea que incluya procesamiento de lenguaje natural, podemos preguntarnos hasta donde el programa tiene conocimiento de gramática o del significado de ciertas palabras o incluso del dominio o contexto de la aplicación, el problema radica en que generalmente el significado se encuentra dentro de las instrucciones que especifican que tarea realizar, es decir que palabras por si solas no tienen ningún significado pero por separado pueden representar distintas cosas, una forma de resolver este problema es representando las reglas y principios que definen el contexto de la aplicación como estructuras simbólicas que pueden ser manipuladas por la aplicación, este método ha sido bastante exitoso en aplicaciones de Inteligencia Artificial, en donde este conocimiento basado en reglas ha permitido que se pueden realizar tareas antes limitadas al razonamiento humano.

La representación de programas mediante gramáticas permite al programador definir el significado de algo de la misma forma que un lenguaje descriptivo, con esta representación la computadora puede ser capaz de generar e interpretar oraciones que cumplan con su gramática.

Pero la capacidad de comprender el significado de una oración no consiste únicamente en saber que algo forma parte de su gramática, si bien la gramática, por si sola que no es mas que una definición abstracta de un conjunto de objetos bien estructurados y que utilizando un *parser*, encargado de interpretar la llegada de cada uno de los objetos que forman la gramática y obtener la estructura que forma dicho lenguaje.

Uno de los principales problemas que cualquier persona que trata de diseñar un sistema que sea capaz de interpretar y relacionar palabras y obtener su significado es la ambigüedad, los lenguajes naturales son propicios para la aparición de ambigüedades casi a todo nivel de descripción, desde el fonético hasta el sociológico, dando como resultado que el *parser* de una gramática que presente ambigüedades pase bastante de su tiempo tratando de investigar la estructura de la oración que presente este fenómeno.

2.2.2 Representación del significado

La comprensión del lenguaje envuelve formas lingüísticas relacionadas del significado, la generación de lenguaje comprende lo contrario, podemos representar formas lingüísticas en las computadoras, pero no su significado, puesto que una computadora únicamente maneja estructuras simbólicas que representan el significado de una oración. Hay varias ideas distintas acerca de la forma en que el significado puede ser representado en humanos y máquinas.

La mas aceptada es la que utiliza la Inteligencia Artificial en sistemas que se basan en reglas, la cual permite considerar al significado como un proceso a realizar, es decir que quien oye será capaz de construir este procedimiento y entonces decir si realizarlo o no.

Esta perspectiva del significado es conocida como ‘semántica procedural’, en la cual el significado del comando es un procedimiento que para ser realizado requiere una acción, por ejemplo, el significado de una pregunta es el de encontrar una respuesta, el significado de una oración es el de agregar nueva información al modelo de quien escucha, pero incluso bajo esta perspectiva resulta difícil determinar cuándo la acción o procedimiento es realizado, o incluso si hay varios procedimiento para un solo comando determinar porque utilizar este y no otro, es decir tomar una decisión, siendo esta una de las principales áreas de investigación de la Inteligencia Artificial.

Investigaciones en como representar el conocimiento han llevado a distintas técnicas para representarlo, desde redes semánticas, siendo una de las más llamativas la de Dependencia Conceptual de Schank, esta fue creada como un modelo psicológico de cómo la gente representa el significado de una oración, los filósofos, Katz y Fodor han propuesto la idea de que partiendo de una serie de primitivas es que el significado es construido. Schank continuó esta tradición proponiendo una serie de 11 primitivas, las cuales combinadas podían representar casi cualquier evento del mundo, la dependencia conceptual fue diseñada independiente del uso de cualquier lenguaje.

La idea de representar el significado del lenguaje natural mediante la lógica es uno de los aspectos que se presenta actualmente en el procesamiento del Lenguaje Natural. El principio de composición también conocido como el principio de Frege dice que el significado de cualquier frase puede ser obtenido por alguna operación en los significados de sus partes, es decir que a partir de una descripción estructural de una oración, es posible obtener el significado de una oración encontrando primero el significado de las palabras en forma individual, primero combinando estos para construir el significado de frases pequeñas, luego combinando estos significados construidos en frases mas largas hasta que el significado de toda la oración es formado.

Otra técnica para representar el conocimiento es el uso de marcadores semánticos y restricciones selectivas como una técnica cruda que puede ser eficiente en términos computacionales, especialmente en dominios restringidos, que presentan cierta ambigüedad. Hasta ahora hemos visto como se puede representar el significado y cómo algunas técnicas permiten resolver la ambigüedad basándose en esta representación del significado.

Pero es necesario para completar un modelo de comprensión de lenguaje, un modelo de conocimiento que permita representarlo y recuperar información de él, y no se puede construir una computadora con una amplia comprensión sin proveerla de un conocimiento casi enciclopédico del mundo.

Si bien estas aseveraciones parecieran pesimistas tienen como intención sugerir que para lograr un objetivo tan amplio como el de obtener una maquina capaz de 'leer' y 'comprender' de una forma más humana, se debe clasificar de forma rigurosa las clases de conocimiento que guían al usuario de un lenguaje y codificar lo suficiente como para saber que nuestro modelo cumple de una forma realista con el trabajo que se desea cumpla.

2.3 Relacionar palabras: Cuan aceptable puede ser la respuesta de un sistema experto para evaluar el sentido de una oración.

Como vimos anteriormente las representaciones del significado son posibles, se pueden usar distintas técnicas desde *scripts* o cuadros como los utilizados por Schanks[3] para sus primitivas, para codificar el significado de documentos completos, o bien utilizar otros métodos de teoría sintáctica mediante estructuras D o LF, redes semánticas, formalismos lógicos e incluso representaciones probabilísticas, en esta parte del capitulo únicamente nos enfocaremos y evaluaremos a aquellas que son adaptaciones de la teoría que trata de proveer del poder predictivo y explicativo de la habilidad lingüística

[3] Ward Church, Kenneth. Murray Hill, Patrick Hanks , Word Association Norms, Mutual Information, and Lexicography Bell Laboratories, CoLlins Publishers, Glasgow, Scotland

humana. Muestras especificas de la interpretación dan como resultado tres puntos específicos concernientes al lenguaje humano[4].

1. Las oraciones significan distintas cosas del mundo, es decir que hay una conexión entre el mundo y el lenguaje.

2. Los humanos hablan y entienden ilimitado número de oraciones.

3. Los humanos pueden hablar y entender mas de un lenguaje humano, mientras que no piensan en ninguno de estos.

Muchos puntos de vista actuales para representar el significado de una oración, se basan en estructuras formales que codifican el significado de estas, pueden ser procesadas sistemáticamente desde una estructura sintáctica de la oración, este método intuitivo tiene raíces en la teoría de la verdad de Tarsk llamada teoría T. Esta teoría tiene las fortalezas y debilidades de varias técnicas de representación del significado tales como las tres que se describen a continuación

1. La estructura del lado izquierdo de la bi condicional no tiene relación necesariamente con la parte derecha.

---

[4] Ward Church, Kenneth. Murray Hill, Patrick Hanks , Word Association Norms, Mutual Information, and Lexicography Bell Laboratories, CoLlins Publishers, Glasgow, Scotland

2. Las bases del sistema en un tipo de relación palabra – mundo son dejadas de algún modo abiertas.

3. La aplicabilidad de la teoría T dependerá de la formación del objeto lenguaje.

A este punto es necesario saber porque es necesaria la teoría T para la representación y recuperación del significado, la razón es que estas teorías dan a menudo objetos - lenguajes exactos que podrían decir que esta o aquella oración significan algo, por ejemplo:

'La nieve es blanca ' es verdad solamente si la nieve es blanca.

convirtiéndolo a una forma más normal

'La nieve es blanca ' significa que la nieve es blanca.

Debido a que estos modelos pueden ser definidos por nosotros, basta con cualquier pequeña variación de la serie de reglas que lo definan para cambiar el significado de una oración. Muchos lingüistas computacionales han decidido no ir mas allá de aquellas restricciones básicas del lenguaje, siguiendo con el enfoque de las teorías T para obtener el significado de una oración, vemos que presenta estos puntos importantes del lenguaje:

- Podemos hablar y entender oraciones simples.
- Podemos hablar y entender un número infinito de oraciones.
- El significado y la estructura de un lenguaje están relacionadas si las oraciones tienen relaciones lógico – semánticas con otras oraciones.

Los primero dos puntos se refieren al carácter generativo del lenguaje. Esto es que el lenguaje es infinito mientras que el significado para comprenderlo es finito.

Estas características del lenguaje fuertemente implican que una teoría para representar el significado debe ser compuesta, de acuerdo al significado de las oraciones basadas en el significado de las partes. El tercer punto es también una fortaleza que presenta la teoría T, si el metalenguaje usado es un formalismo sobre el cual han sido definidos procedimientos de inferencia, el razonamiento automatizado puede ser precedido por el lado derecho de la regla en una bi condicional (en este caso la oración).

Aunque las teorías T también poseen ciertas debilidades, dejando ciertos aspectos sin cubrir, que son dejados para otro nivel de análisis, por ejemplo:

- Las oraciones significan cosas, es decir que relacionan hechos con el mundo.

- Las oraciones son algunas veces ambiguas.

- Es fácil construir oraciones anómalas, que siguen siendo oraciones.

No obstante ninguna de estas limitaciones hace que las teorías T dejen de ser bastante utilizadas en sistemas de información, a pesar de que estas debilidades harán que incluso oraciones simples puedan no ser interpretadas por este método.

Mediante el desarrollo de una Estructura Conceptual Léxica, es posible resolver algunas de las áreas que las teorías T dejan libres. La mayoría de las teorías interpretan a los verbos como siendo parte de la estructura del predicado, Jackendoff, creador de la Estructura Conceptual Léxica o LCS trata de profundizar en esta idea con el análisis sintáctico. El objetivo de su análisis es que uno puede caracterizar los argumentos semánticos hacia el verbo de forma sintáctica, habiendo para ellos dos principios envueltos en esta tarea.

2.3.1 Otras teorías de análisis semántico

2.3.1.1 La teoría de la barra X

Esta restringe la forma en que los árboles describen la estructura interna de una oración. Los artículos léxicos de cualquier categoría mayor, pronombres, verbos adjetivos, adverbios, o preposiciones pueden iniciar frases con sus correspondientes tipos, iniciar una frase significa que la estructura que proyecta al articulo léxico, es la que maximiza a sus modificadores.

2.3.1.2 La teoría θ

De acuerdo a esta teoría, parte de lo que significa conocer un verbo es el conocer que representa.

El verbo puede asignar distintos roles (al sujeto), siendo ninguno de estos obligatorio, pero si aparecen en la oración es que ellos(los roles) son argumentos de los verbos. Con estas dos teorías es posible obtener estructuras semánticas preliminares de estructuras sintácticas, sabiendo que debemos buscar las posiciones del sujeto y el complemento, para frases que llenen estos roles y que pueden ser tratadas en otras posiciones del texto de forma diferente (debido a que ninguna transformación pasiva ha sido operada)

2.4 Principios del reconocimiento inteligente de caracteres

El término de Reconocimiento Inteligente de Caracteres (ICR), agrupa a varias tecnologías que apuntan a su vez al análisis y reconocimiento de caracteres escritos a mano a partir de imágenes electrónicas. El problema puede ser definido así: dada una imagen digitalizada, como podría un algoritmo ICR analizar su contenido, reconocer la identidad de varios caracteres contenidos en la imagen y ser capaz de devolver esa información.

Una imagen digitalizada es después de todo una colección de números, para imágenes binarias a cada punto o píxel le es asignado un valor ya sea 0 o 1, siendo este el punto de partida para los métodos estadísticos que permiten analizar un carácter, estos generalmente funcionan de manera aceptable, pero en el caso de imágenes de documentos en los que aparecen caracteres escritos a mano, estos resultan inadecuados.

Existe otra aproximación a este problema, en la cual, una imagen si bien esta compuesta por píxeles, no solo es solamente esto, sino que también estos píxeles formaran el contorno de una carácter escrito a mano, lo cual hará que este método sea más aplicable a un documento que contenga este tipo de caracteres.

Al seguir el contorno de un carácter, punto por punto, es posible determinar cual es, y en algunos casos, este método proveerá soluciones similares a caracteres con características de escritura similares, siendo esta su fortaleza y debilidad. Al mezclar estos métodos es posible crear un sistema híbrido que aplica la fortaleza de ambos, mientras reduce sus debilidades; actualmente es posible encontrar distintos tipos de motores, la unión de varios de los algoritmos de estos motores permiten crear un algoritmo en el cual se elige la solución basado en la respuesta optimizada de estos. Hasta este momento se han mostrado distintas metodologías que buscan mejorar el rendimiento de un motor OCR a fin de darle comprensión sobre lo que interpreta de una imagen digitalizada, pero puede entonces surgir una pregunta ¿Para que puede servir el que una computadora pueda asimilar algo que puede 'leer'?

2.5 Utilizando técnicas híbridas de OCR para romper *CAPTCHA´s*

En palabras del autor del término *CAPTCHA* L. von Ahn, *"Cualquier programa que pase una prueba generada por un CAPTCHA, puede ser utilizado para resolver cualquier problema no resuelto de la inteligencia artificial"*.

Actualmente existen varios grupos de investigadores que se han dado a la tarea de completar algoritmos de reconocimiento óptico que permitan romper algunos de los *CAPTCHA* actuales, logrando uno de estos equipos porcentajes de exactitud de un 93% para uno de los CAPTCHA utilizados por YAHOO para la apertura de cuentas de correo electrónico, el EZ – Gimpy, en esta sección se describen las técnicas utilizadas por este equipo de investigadores.

Es decir que si un problema de inteligencia artificial es resuelto por alguno de los actuales algoritmos OCR puede ser que uno de estos mismo algoritmos resuelvan cualquier otro problema de la Inteligencia Artificial (robotica, procesamiento de lenguaje natural, etc.)

2.5.1 ¿Que es un *CAPTCHA?*

La palabra *CAPTCHA* es el acrónimo en ingles para “Prueba de Turing completamente automatizada para diferenciar computadoras y humanos” (en ingles, *Completely Automated Public Turing test to tell Computers and Humans Apart*), el cual consiste en una prueba desafío – respuesta, utilizada para determinar si efectivamente el usuario es humano o no. El termino fue acuñado en el año 2000 por Luis von Ahn, Manuel Blum y Nicholas J. Hopper de la Carnegie Mellon University, y John Langford de IBM.

2.5.1.1 La prueba de Turing (*Turing Test)*

La prueba de Turing es una propuesta para probar la capacidad que tiene una máquina para realizar conversaciones parecidas a las de los humanos. Descrita por Alan Turing en el articulo *Computing machinery and intelligence*, basada en un juego común en las fiestas en las que los invitados intentaban descubrir el sexo de la persona en otra habitación utilizando para ello mensajes escritos, la prueba de Turing funciona de la siguiente forma: un juez humano entabla una conversación en lenguaje natural con otras dos entes, una humana y la otra una maquina, si el juez es incapaz de identificar correctamente cual es cual, entonces se dice que la maquina paso la prueba, se asume que tanto la maquina como el humano pretenden actuar de forma "humana". En orden de mantener simple y universal esta prueba, la conversación ha sido usualmente limitada a canales de texto únicamente.

Originalmente Turing propuso esta prueba para reemplazar la que el consideraba una pregunta sin sentido "¿Pueden las maquinas pensar?" por una mucho mas adecuada, prediciendo a su vez que eventualmente las maquinas serian capaces de pasar la prueba, en efecto el predijo que para el año 2000 las computadoras con unos 120 Mb de memoria serian capaces de engañar a un 30% de los jueces humanos durante una prueba de cinco minutos.

Aunque hay quienes aseguran que la prueba de Turing no podria servir para definir de forma valida la inteligencia de una máquina pensante, por al menos tres razones:

1. Una maquina que pase la prueba de Turing, puede que sea capaz de simular el comportamiento de un humano, pero esto seguiria siendo mucho mas debil que la verdadera inteligencia. La maquina puede que unicamente siga algunas reglas inteligentemente definidas (aunque incluso los humanos seguimos algunas reglas definidas de forma inteligente).

2. Una maquina puede muy bien ser inteligente sin ser capaz de hablar como lo haría un humano.

3. Muchos humanos que podrían ser considerados inteligentes podrían fallar en esta prueba (por ejemplo alguno de corta edad o sin educación)

Dependiendo del tipo del *CAPTCHA*, este solicitara al usuario que escriba las letras de una imagen distorsionada y obscurecida de una secuencia de letras o dígitos que aparecen en la pantalla

Figura 5. Imagen de EZ Gimpy Captcha

Este CAPTCHA de las secuencia de letras “smwm” dificulta la interpretación por computadora, distorsionando la imagen mediante efectos que retuercen los bordes de las letras, y agregando un fondo de color generado por gradientes

2.5.2 Aplicaciones

Los CAPTCHA son empleados para prevenir que los BOT(derivado de la palabra ROBOT, identifica en el Internet a un programa que funciona como una agente de software que interactúa con otros servicios de la red, para que funcione como si fuese en realidad una persona, uno de sus usos mas comunes es el de reunir información) utilicen varios tipos de servicios computacionales.

Sus aplicaciones van desde prevenir que los BOT tomen parte en encuestas en línea hasta evitar que se puedan registrar en cuentas de correo electrónico gratuito (que después puede ser utilizado para enviar correo no solicitado o *spam*), y mas recientemente evitar que correo *spam* generado por un *BOT* llegue a su destinatario, obligando al que lo envía a pasar una prueba de CAPTCHA.

2.5.3 Características

Por definición, los CAPTCHA tienen las siguientes características:

- La mayoría de los humanos puede pasar una de estas pruebas pero los actuales programas de computadora, incluidos los algoritmos OCR, NO PUEDEN pasar.
- Son completamente automatizados, esto evita la necesidad de mantenimiento o intervención humana en la prueba, con los obvios beneficios de costo y confiabilidad.
- El algoritmo utilizado para crear el CAPTCHA es publico, aunque protegido por alguna patente, esto es así debido a que el romper una CAPTCHA conlleva gran trabajo en el área de la inteligencia artificial, mas que el hecho de descubrir el algoritmo con el que fue hecho, que puede ser encontrado mediante ingeniería inversa u otros medios.

3. RECONOCIMIENTO ÓPTICO DE CARACTERES VRS. EL MÉTODO TRADICIONAL DE PROCESAMIENTO DE INFORMACIÓN.

3.1 El método tradicional

El método mas comúnmente empleado para el ingreso de información es aquel en el que se ven involucrados operadores encargados de digitar la información recopilada. En el caso de los digitadores, ellos generalmente ingresan artículos, números u otro tipo de datos a las computadoras o bien, completan formas que aparecen en las pantallas. También se encargan de manipular la información existente, editar la información actual o revisar las entradas de nueva información a una base de datos. Algunos ejemplos de la información que un digitador procesa son la información de personal y clientes, registros médicos, membresías, etc. Usualmente esta información es usada internamente por una compañía.

Los digitadores utilizan distintos tipos de equipo para ingresar información. Muchos digitadores utilizan máquinas que convierten la información que ellos escriben a impulsos magnéticos en cintas o discos para ser ingresados a otras máquinas principales; algunos de ellos operan máquinas en línea o en computadoras personales. Para el ingreso de la información se diferencia dos metodologías que son las mas ampliamente difundidas.

3.1.1 Digitación vista hacia abajo (*Heads-Down Keying*)

Esta metodología es común en aquellos casos en el que el volumen de información es bastante grande, por ejemplo en un censo. Mientras se ingresa la información, el operador generalmente no ve hacia la pantalla de su computadora, en cambio, ve hacia abajo a la mesa o superficie de trabajo donde se encuentra el documento que el ingresa.

El objetivo de esta metodología es de transcribir a la computadora tan pronto como sea posible la información que aparece en el documento, la revisión es generalmente mantenida al mínimo para ser resuelta en una fase posterior, generalmente a través de programas que tienen cierta similitud a los algoritmos empleado en OCR o de programas de edición. Los operadores no necesitan estar familiarizados con el contenido del documento, ellos hacen pocas decisiones para resolver los errores en la información. Su más importante habilidad es rapidez y exactitud.

3.1.2 Digitación con la vista hacia arriba (*Heads-Up Keying*)

Esta metodología es empleada cuando la cantidad de información es mas reducida y la complejidad de la información más grande; mientras se ingresan los datos el operador a menudo se sirve de la pantalla de la computadora como una referencia tanto como del documento que transcribe. El objetivo es capturar y corregir tantos errores como sean posibles al ingresar la información.

Los operadores necesitan estar familiarizados con el contenido del documento para así poder tomar decisiones al resolver errores de la información y a la vez ser entrenados para hacerlo.

En los últimos años se ha incrementado una forma de trabajo en la que los digitadores operan formularios sin necesidad de teclados, tales como *scanners* y archivos transmitidos electrónicamente, son cuando se utilizan estos nuevos sistemas de reconocimiento de caracteres que los digitadores intervienen corrigiendo únicamente aquellos símbolos que no pudieron ser reconocidos por las máquinas.

3.2 El procesamiento óptico como una forma alternativa

Durante los últimos años, la tecnología de reconocimiento óptico de caracteres y el reconocimiento óptico de marcas han comenzado a ser vistas como una forma alternativa para capturar y procesar grandes volúmenes de información, siendo esta experiencia para algunos satisfactoria y para otros no tanto, pero para ambos casos se han tenido que enfrentar altas restricciones operacionales. Factores como la calidad del papel del documento origen, la calidad de impresión han sido claves para el éxito de esta tecnología. Un requerimiento básico al realizar operaciones de captura de datos de forma masiva es el de poder confiar en la información que se obtenga de estas operaciones, las nuevas tecnologías permiten acelerar el proceso de ingreso de información y a la vez incrementar la exactitud.

En el caso de los *scanners* ópticos, algunos dispositivos son capaces de leer entre 4,000 y 6,000 páginas por hora. En el caso de este tipo de captura de datos un requerimiento clave al utilizar *scanners* es el de estar seguro que todos los documentos origen sean leídos de manera correcta, y que no hay caracteres adicionales o faltantes, puede ocurrir que la información impresa presente alguna imperfección como el corrimiento de la tinta o que esta sea muy pálida (en el caso de información escrita a mano) y que el *scanner* no haya sido capaz de detectar el carácter correctamente, esta clase de errores son típicos de documentos capturados mediante estos métodos.

3.2.1 ¿Qué costos implica el uso de OCR y cual es el nivel de intervención humana al aplicar esta tecnología?

Para cualquier motor OCR, es necesario evaluar cuanto tiempo le llevará a un operador humano el corregir el resultado de la interpretación OCR de un campo dado. Primero consideramos la automatización de este proceso, utilizando métodos basados en la distancia que hay de editar desde la fuente original al resultado OCR. No existe un método exacto para definir tal distancia, para esto es necesario definir el peso de una sustitución, una eliminación, o una inserción y entonces aplicar un algoritmo de comparación de cadenas, el cual permitirá aproximar el número mínimo de operaciones de edición requerida para corregir el resultado. Este número mínimo puede ser utilizado como una medida del tiempo de corrección del operador.

3.2.2 La intervención humana

Dependiendo del caso, la dificultad es que el operador puede escoger entre re-escribir el campo completo antes que tratar de modificar el campo OCR. Por lo tanto mas que tratar de modelar el proceso, se escoge medir el nivel de dificultad que representa para el operador el editar el campo. Un algoritmo de comparación de cadenas solo logrará expresar las distancias existentes entre ambas cadenas (la original y la resultado) letra por letra, con un criterio de inserción – eliminación – sustitución y no logrará reflejar el costo real que un operador tendrá al corregir los resultados.

El criterio de inserción – eliminación – sustitución ha sido clasificado en 4 clases principales, de acuerdo al esfuerzo realizado por una persona, estas cuatro clases son definidas por reglas y pueden resumirse de la siguiente manera:

- Ninguno (el resultado es suficiente como para ser utilizado)
- Poco (bajos esfuerzos, pocas palabras deben ser corregidas)
- Medio (uno o dos caracteres deben ser corregidos para identificar una palabra).
- Alto (algunas veces es mejor rescribir las palabras)

Siendo este criterio el mas ampliamente utilizado al evaluar motores OCR comerciales, pueden utilizarse distintos costos para cada una de las operaciones que implica (inserción, eliminación, sustitución). El costo de la sustitución es usualmente el mas bajo, debido a que a menudo es directamente aplicable a un error OCR. En muchas aplicaciones, las inserciones y las eliminaciones son raras, siendo causadas por equivocaciones del escritor o problemas en la segmentación del documento. Si el costo de estas operaciones es alto, serán incluidas únicamente en aquellos que tengan grandes probabilidades de ocurrencia.

3.2.3 Facilitar la intervención y la corrección

Para los costos de sustituciones, estos podrán ser estimados a partir de dos fuentes de información: la matriz de confusión del clasificador y una posterior clase condicional de probabilidades, suministrada como salida del clasificador. La primera provee una estimación bastante consistente (debido a que se deriva de una muestra bastante amplia) de una probabilidad previa de un carácter dado mezclado con otro por el clasificador.

La segunda puede ser considerada menos confiable, pero toma en cuenta las presentaciones del carácter bajo análisis, y por lo tanto acuerda una información mas dinámica, comparada con la naturaleza estática de la matriz de confusión.

Al hablar de la automatización del proceso de corrección de salidas OCR se ha de tomar en cuenta que la post – corrección de los resultados OCR para documentos de texto está basada usualmente en diccionarios electrónicos. Cuando se escanea texto de un área temática específica, los diccionarios convencionales a menudo pierden considerable número de *tokens* e incluso si la frecuencia de las palabras es almacenada estas no reflejarán apropiadamente las frecuencias halladas en el área temática dada.

Se han realizado experimentos para comparar el uso de distintos tipos de diccionario, utilizando diccionarios estáticos amplios, el uso de diccionarios dinámicos obtenidos vía un análisis automatizado de páginas Web de un dominio especificado (relacionados con el área temática) y diccionarios mixtos.

Estos experimentos demostraron que el uso de diccionarios dinámicos mejoraron la cobertura del área temática de una forma significativa y ayudaron a mejorar la calidad de los métodos de post – corrección léxica, en el siguiente capítulo se hará una evaluación de un caso real bajo ambas aproximaciones (método tradicional y OCR).

3.3 Procesamiento de Documentos mediante Reconocimiento Inteligente de Caracteres.

3.3.1 Funciones y Características de un sistema ideal

1. Abierto y Escalable: el sistema debería ser completamente escalable, desde una maquina solitaria (*stand alone*) a una configuración de red amplia, utilizando hardware y software estándar, el sistema ICR permitirá al usuario adquirir aquel software y hardware que necesite para sus actuales necesidades y conforme el sistema incremente su capacidad, según la carga de trabajo que se tenga. En procesamiento a gran escala, la configuración opera en un ambiente de red con múltiples *scanners* conectados a uno o varios servidores.

2. Flexibilidad (Múltiples Motores de Reconocimiento): el sistema opcional deberá primeramente incorporar la mejor tecnología de reconocimiento de caracteres para poderlo transferir al servidor de información, donde pueda ser procesado por varios motores de reconocimiento antes de proporcionar una salida, además estos deberán proveer otras herramientas entre ellas OCR / ICR, identificadores de marcas, código de barras y la capacidad de definir reglas y especificaciones de salida.

3. Soportar alimentación electrónica (Fax, E Mail, Internet): Los mejores sistemas deben trabajar con imágenes, sin importar su fuente, pero si su calidad, los motores, el “ruido” que presente una imagen puede ser determinante en la calidad de la información que logre capturar.

4. Funcionalidad en archivos (corto y largo plazo, contingencias) otra de las características de un sistema de este tipo es que sea capaz de adquirir imágenes de cualquier otro producto de similares características. Además de esto el usuario podrá almacenar imágenes para un uso posterior, incluso siendo estas empaquetadas para reducir la cantidad de espacio que ocupen.

5. Compatibilidad con varios sistemas operativos.

6. Capacidad de exportación a formatos estándar.

7. Permitir la configuración de múltiples interfaces de *scanner* que permitan variar las entradas de la información.

8. Deberá poder transferir información, de ser necesario a otras estaciones en otras redes.

9. Soportar otros procesos: como un sub sistema de captura, los sistema ICR pueden capturar, verificar, y proveer datos e imágenes a cualquier otro proceso capaz de importar información en formato ASCII o imagenes.

10. Un sistema ICR ideal debe poder ser controlado totalmente por los usuarios, una funcionalidad significativa deberá ser ofrecida para permitir a los usuarios definir plantillas de los formularios, reglas, parámetros, y cargas de trabajo.

11. Capacidad para identificar aquellas salidas del procesamiento que no coincidan con ninguna palabra o bien que posean datos irreconocibles.

12. Capacidad “Vista hacia Abajo”: la tecnología ICR permite al operador digitar a partir de una imagen (*Key From Image* o KFI por sus siglas en ingles) toda aquella información que no pudo ser capturada por reconocimiento de caracteres o marcas. Estudios han demostrado que al procesar información mediante KFI, esta metodología resulto ser mucho mas efectiva que al hacerlo mediante digitación “Vista hacia abajo”. Aun así, es posible configurar un sistema de este tipo para que ciertos campos de un formulario puedan ser digitados mediante la metodología de digitación “vista hacia abajo”.

13. Funcionalidad de codificación: utilizando conexión a bases de datos, la información puede ser comparada y enviada de una forma apropiada.

14. Exactitud de Reconocimiento: la exactitud en un sistema ICR ha mejorado en los últimos años, debido a la combinación de mejoras en las redes neuronales, motores, *scanners*, interfaces y herramientas de desarrollo para estas, aun así, es quiza el aspecto mas incomprendido y a la vez determinante al escoger un sistema de este tipo para procesar información.

3.3.2 Consideraciones al evaluar el rendimiento de un sistema ICR / OCR

La mayoría de personas encargadas de evaluar tecnología ICR se han enfocado en su exactitud, y más específicamente en las tasas de aceptación, es decir aquella medida de errores a partir de aquellos caracteres que no fueron leídos y mostrados para ser editados.

Mientras que esto es una forma de medir el rendimiento del sistema, esto generalmente hace que se ignore el problema mas critico al evaluar este tipo de tecnología, siendo este el aumento de la productividad al procesar información, mas que la exactitud o tasas de aceptación.

La productividad de un sistema ICR se basa en una función de varias variables, todas ellas pueden ser controladas a distintos niveles, estos variables incluyen:

- Restricciones en el diseño de formularios (tamaño, forma, cuadros de texto bien delimitados) en comparación de cuadros de texto con pocas o ninguna restricción

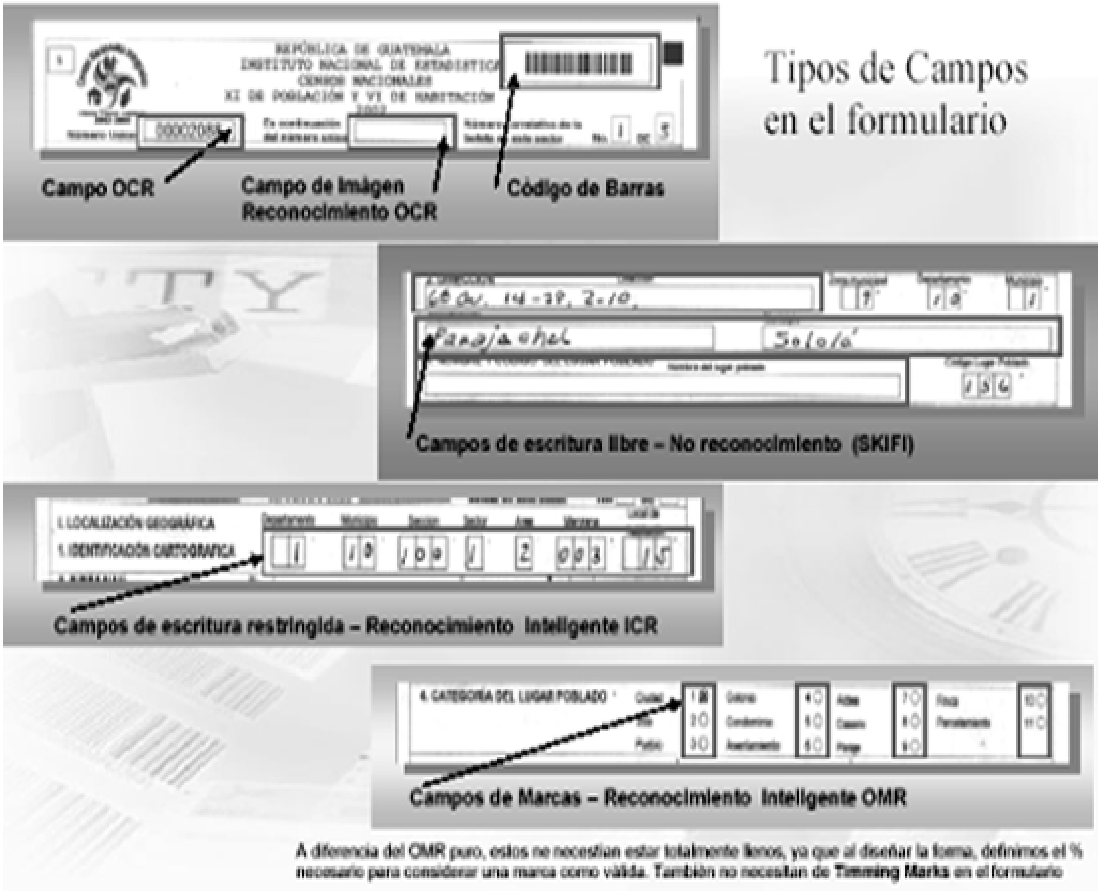
- Los colores de extracción (color de la escritura, fondo del formulario), la . remoción de líneas que delimiten el texto, la forma de escritura a mano, marcas o impresión tipográficas, son consideraciones criticas, si

estos factores son tomados en cuenta, el reconocimiento podrá mejorar de forma dramática.

- La habilidad en desarrollar y afinar la aplicación y sus herramientas pueden mejorar los resultados. Por ejemplo, se puede configurar un sistema de manera que este elimine aquellos caracteres que no reconozca, pero también puede que el resultado se vea seriamente afectado por esto. El diseño de formularios y el afinamiento de la aplicación debe ser idealmente hecho por el mismo personal puesto que los resultados de la aplicación podrán mejorar si el afinamiento es el adecuado.

- El nivel de escolaridad de la persona que llene el formulario, si recibió algún entrenamiento para llenar los campos de un formulario, se podrá reducir la tasa de errores de los algoritmos al no reconocer algún símbolo
- Definir los límites de un campo puede reducir los errores, al limitar a la persona que llena el formulario a elegir respuestas pre definidas, que bien pueden ser representadas por marcas.

Figura 6. Cuadros de texto en un formulario ya digitalizado para su posterior reconocimiento



4. EVALUACIÓN DEL RENDIMIENTO DEL SOFTWARE DE RECONOCIMIENTO ÓPTICO

4.1 ¿Cómo funciona el software OCR / ICR?

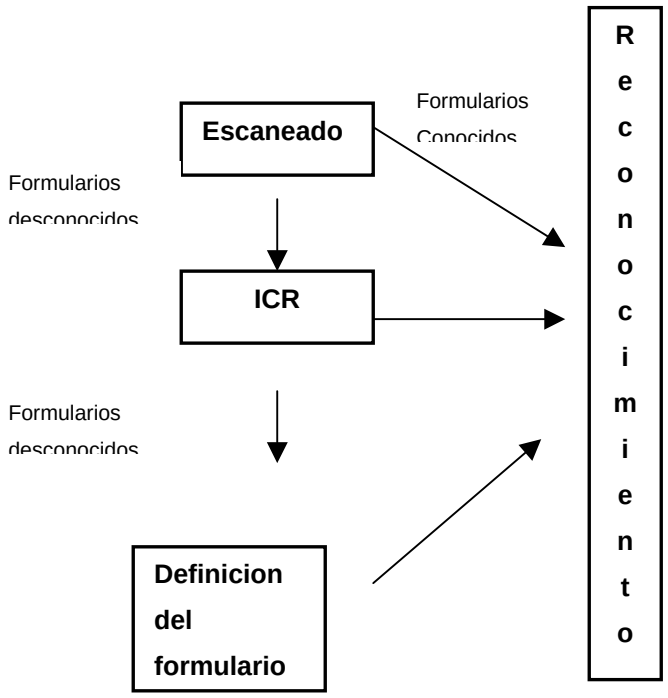
Hasta este punto se ha mostrado como se procesa una imagen para obtener de ella un carácter y de ahí al unirlo como se forman palabras, se ha observado y explicado además algunos de los algoritmos que permiten a los usuarios de distintos sistemas de procesamiento, el ver en la pantalla de su computadora la salida corregida que un motor OCR provee al alimentarlo con una imagen electrónica de un documento obtenida electrónicamente o de un documento escaneado.

Todos los usuarios de computadoras personales alrededor del mundo han tenido en determinado momento algún tipo de experiencia con un motor OCR, el mas común que actualmente se haya en casi cualquier computadora del mundo occidental es el que se utiliza en el *Acrobat Reader* comercializado por Adobe Inc., este motor es capaz de reconocer el texto provenientes de imágenes almacenadas en archivos de tipo PDF, para su exportación hacia cualquier otro editor de texto, si bien este motor reconoce únicamente caracteres tipográficos, sirve de ejemplo para mostrar al lector que tan común se ha vuelto el uso de la tecnología OCR en la vida diaria.

4.1.1 Análisis de las características básicas de software OCR/ICR

Si bien todo software debe tener características que lo hagan ser reconocido como un producto de la buena ingeniería del software, en el caso de un motor OCR este debe de poseer características adicionales que le permitan, debido a su naturaleza el poder analizar soluciones simultaneas que se hagan en un tiempo de respuesta corto o por lo menos aceptable.

Figura 7. Componentes básicos de software capaz de interpretar formularios que contengan caracteres escritos a mano o marcas.



4.1.2 Software OCR comercial aplicados al procesamiento masivo de documentos

Alrededor del mundo se han desarrollado distintos motores OCR que permiten procesar grandes cantidades de información, a gran velocidad y exactitud, utilizando para ello escáneres de alta velocidad y resolución, a la vez de utilizar algoritmos mejorados de adelgazamiento que permiten identificar un carácter.

Tabla II. Software disponible para el reconocimiento óptico de documentos

Fuente	Paquete(s)	Sitio Web	Notas
AABBYY	FineReader	http://www.abbyy.com	Comercial
Scansoft	OmniPage OmniForm TextBridge	http://www,scansoft.com	Comercial
ExperVision	TypeReader	http://www.expervision.com	Comercial, OCR en linea, servicio gratuito
CharacTell	Simple OCR	http://www.simpleocr.com	Comercial
Musitek	SmartScore	http://www.musitek.com	Comercial
Michael D.	Form-Based	http://www.itl.nist.gov/	Gratis, con codigo fuente
Garris (NIST)	HandPrint	Iaui/894.03/databases/defs/nist_ocr.html	Uso sin restricciones, nivel de adiestramiento alto
Donato Malerba	Wisdom++	http://www.di.uniba.it/~malerba/wisdom++/	Gratis para investigacion y propósitos de enseñanza
R. Karpischek	Clara OCR	http://www.claraocr.org	GPL+
J. Schulenburg	JOCR	http://jocr.sourceforge.net	GPL+
Klaas Freitag	Kooka y OCR Interface	http://www.kde.org/apps/kooka	GPL+, Escaneo

Fuente: A Survey of Table Recognition: Models,Observations, Transformations, and Inferences, R. Zanibbi, D. Blostein and J.R. Cordy

4.1.2.1 FormsPro

Desarrollado por Oyster Software, Forms Pro es una herramienta que permite reemplazar el ingreso de información almacenada en papel, a un proceso mas automatizado, que permite procesar documentos basados en formas utilizando tecnologías OCR / ICR (*Intelligent Character Recognition*) e IFR (*Intelligent Form Recognition*). Para que FormsPro pueda funcionar es necesario que el usuario defina los aspectos lógicos y gráficos de la forma, utilizando la herramienta de definición de formularios, cumpliendo con uno de los requerimientos básicos para un software de este tipo.

La función de definición de formularios provee un único sistema integrado en el que la configuración del *scanner*, los parámetros OCR / ICR, las rutas del flujo de trabajo (*Workflow*), las reglas de validación y la edición de la información puede ser hecha por el usuario, sin ningún tipo de programación. El usuario define el formulario, escaneando una muestra de el, es entonces cuando el usuario define los campos y posiciones que se requieren de la imagen escaneada, especificando el nombre, el tipo de dato y su longitud. (Ver figura 10)

Luego de especificadas las características del formulario, se procede a el escaneo de los documentos, a la vez que permite al usuario realizar correcciones sobre la configuración.

Luego de escanear, la imagen es procesada por la aplicacion de reconocimiento, la cual ocurre en dos fases, utilizando IFR para identificar el formulario y pre procesarlo. Esto incluye analizar el formulario para localizar marcas, regiones y campos en el formulario.

Figura 8. Escaneo para la configuración y selección de campos en FormsPro.

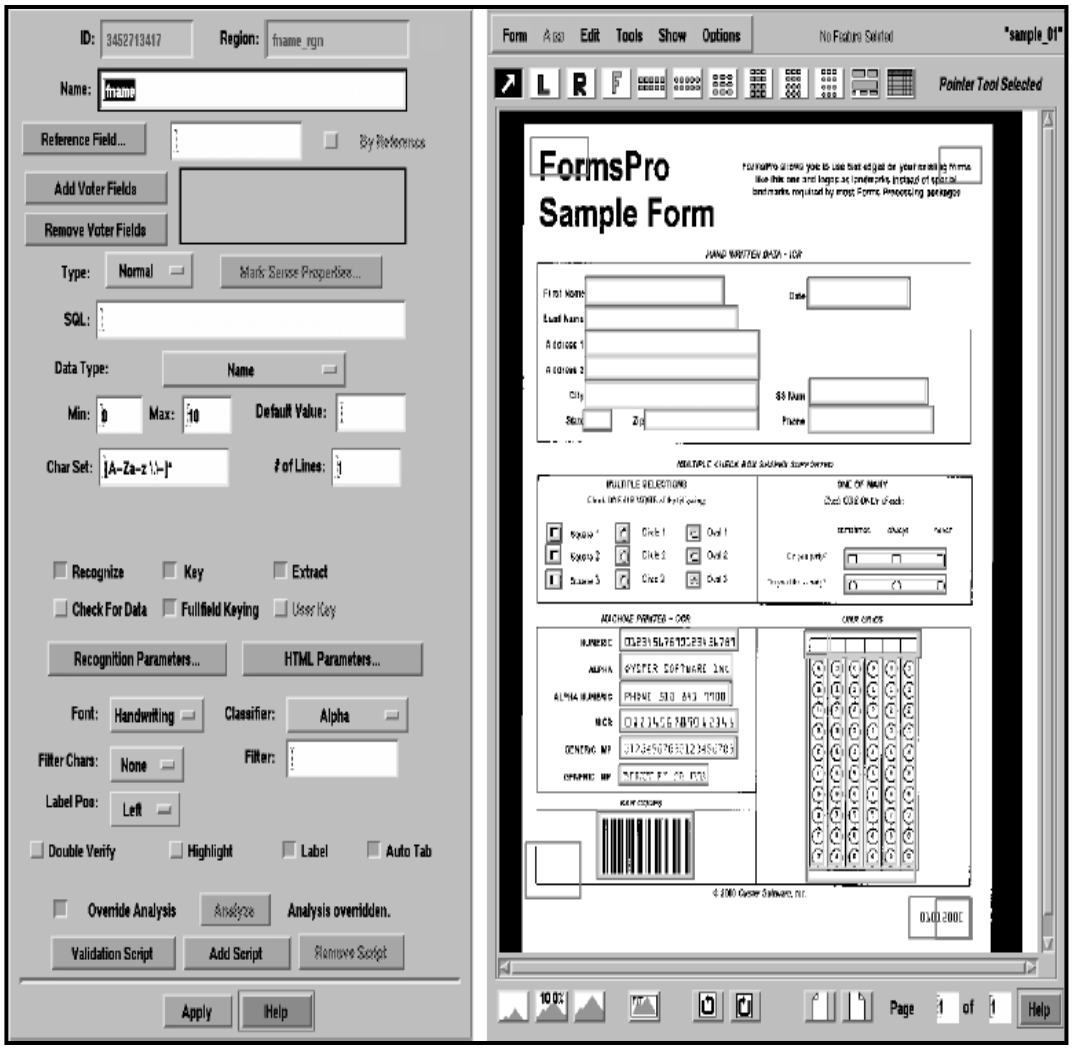


Figura 9. Escaneo en bloque para el formulario anteriormente definido

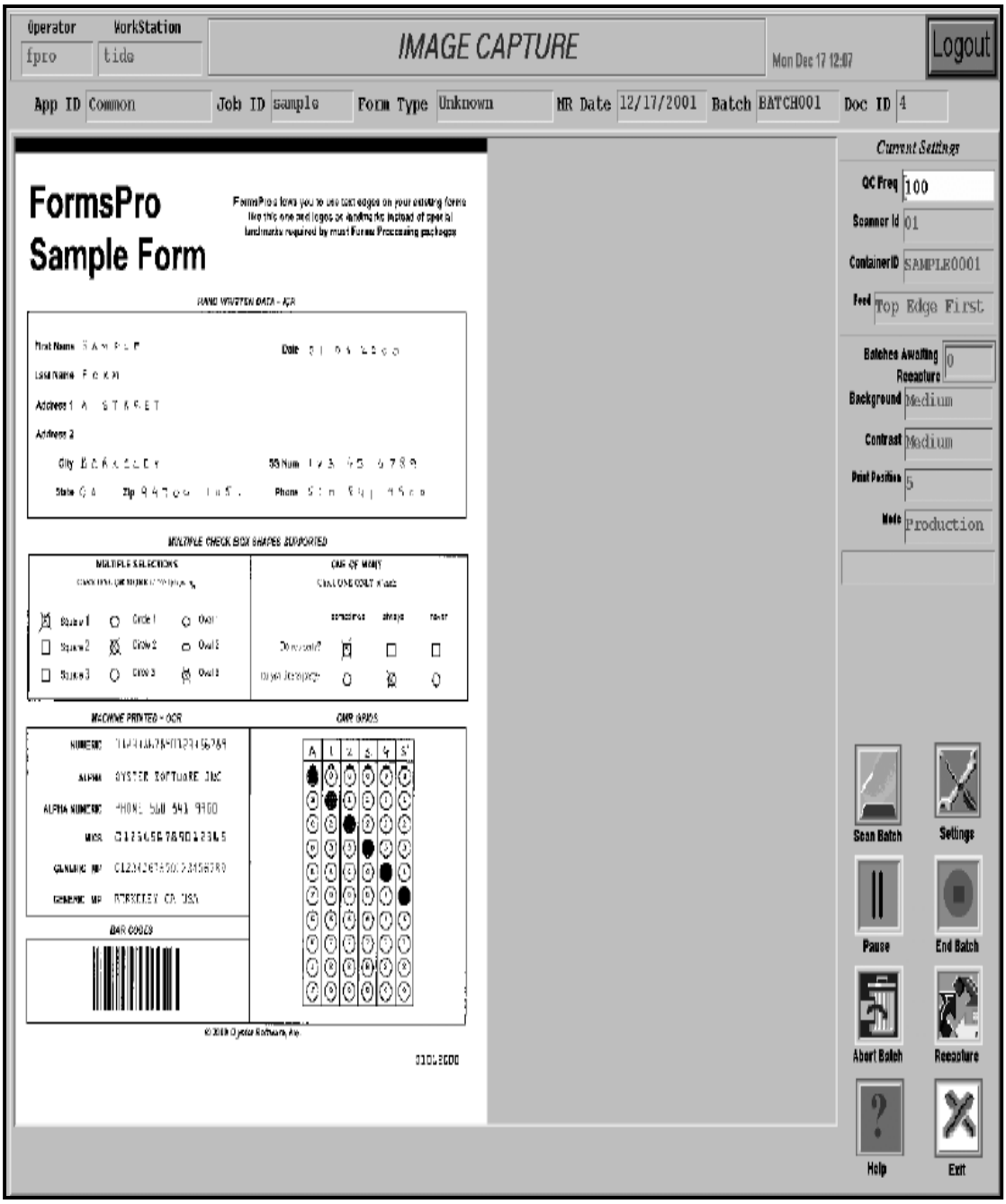
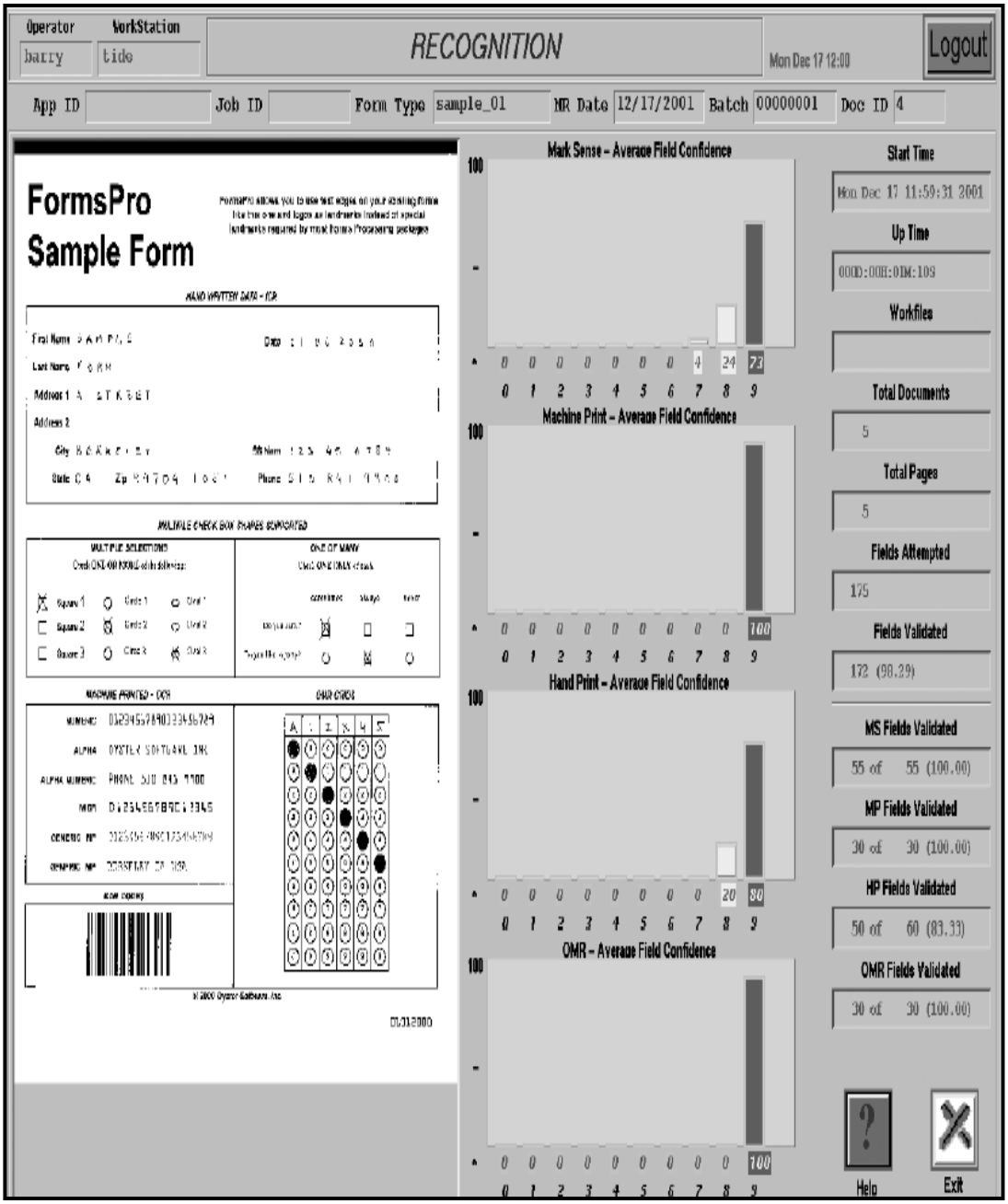


Figura 10. Fase de reconocimiento, en la que el software muestra todos aquellos campos definidos en el y que al procesar la imagen ha podido identificar satisfactoriamente.



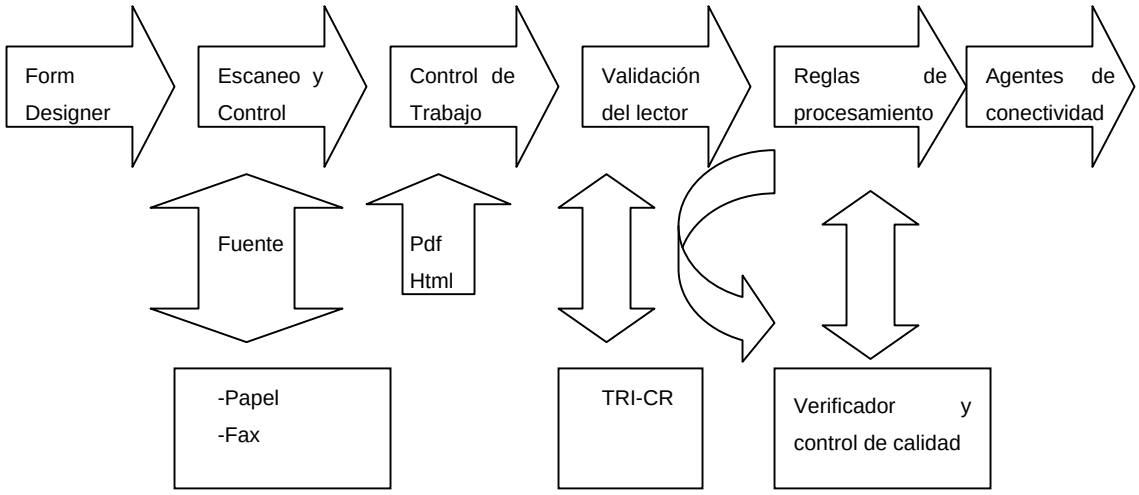
4.1.2.2 TeleForm

Desarrollado por la empresa Cardiff, es utilizado por negocios y organizaciones de gobierno para tener un sistema rápido y fácil de captura y conversión de formas en papel creadas por el mismo software a datos digitales que pueden ser utilizados y explotados. Teleform ofrece una solución para reducir los costos de operación asociados a digitar todos los datos manualmente, estimando que disminuye el tiempo de respuesta, incrementa calidad del dato y acelera la presentación de los resultados.

Este software utiliza el concepto de *batch* (lote), lo cual es un conjunto de cuestionarios de campo con la misma identificación, los cuales examina a través de un escáner y luego almacena la imagen. Una vez almacenada la información pasa por el proceso de reconocimiento que utiliza 5 motores los cuales son como un sistema de elección para determinar que carácter es el que se esta reconociendo.

Si tres de los electores reconocen el carácter, es aceptado; de lo contrario se almacena en repositorio para ser verificado por personas que hacen de digitadores. A ellos se les activa un menú con los batch pendientes de revisar y muestra las imágenes (cuestionarios escaneados) y pide la confirmación de los carácter con duda.

Figura 11. Arquitectura de Reconocimiento y captura de TELEFORM



Dentro de Teleform se encuentran los siguientes módulos:

- Creador de formas: con este modulo se diseñan las boletas que serán utilizadas para la recolección de datos en campo. Además digitaliza y configura formas existentes, define características y especificaciones de las formas, define opciones de exportación y verificación de los dataos. Aquí se pueden escribir *scripts* de programas en Visual Basic y otras herramientas de desarrollo. Es importante mencionar que al definir la forma se pueden hacer enlaces con tablas descriptivas de códigos, establecer rangos permitidos y hacer codificaciones de preguntas abiertas de manera asistida.

Figura 12. Boleta Censal creada en TELEFORM

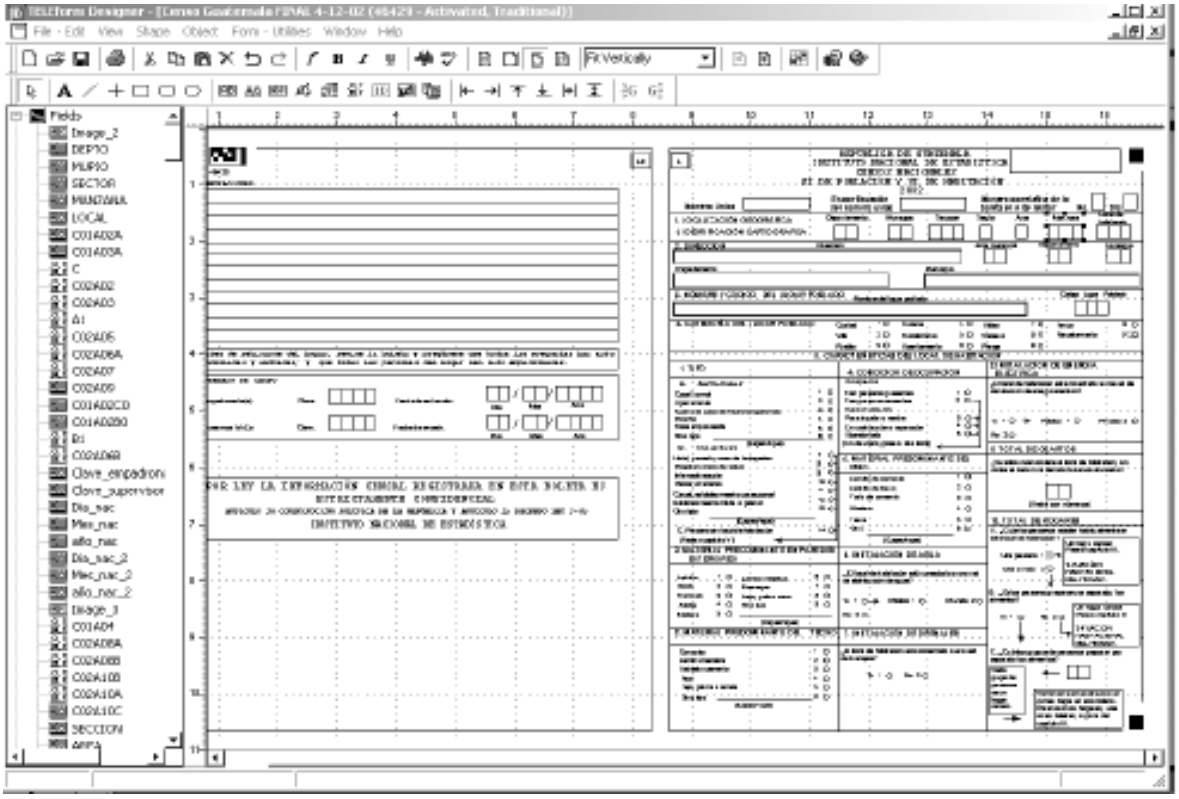


Figura 13. Areas que se desea sean reconocidas en una boleta, los cuadros de texto enmarcados ejemplifican aquellas regiones de las boletas de las que se desea recuperar información

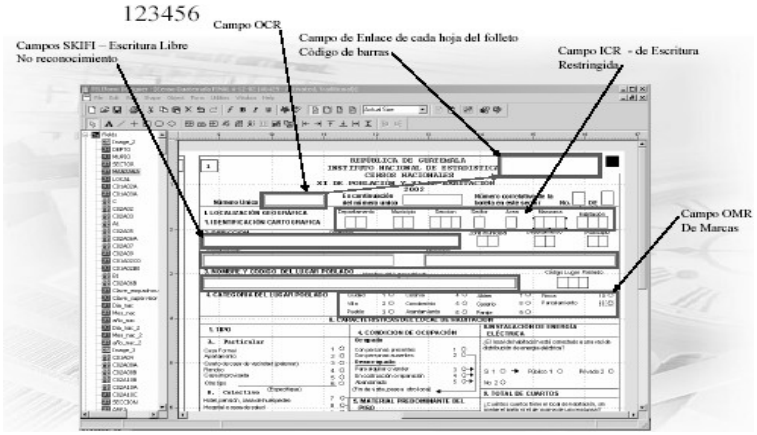
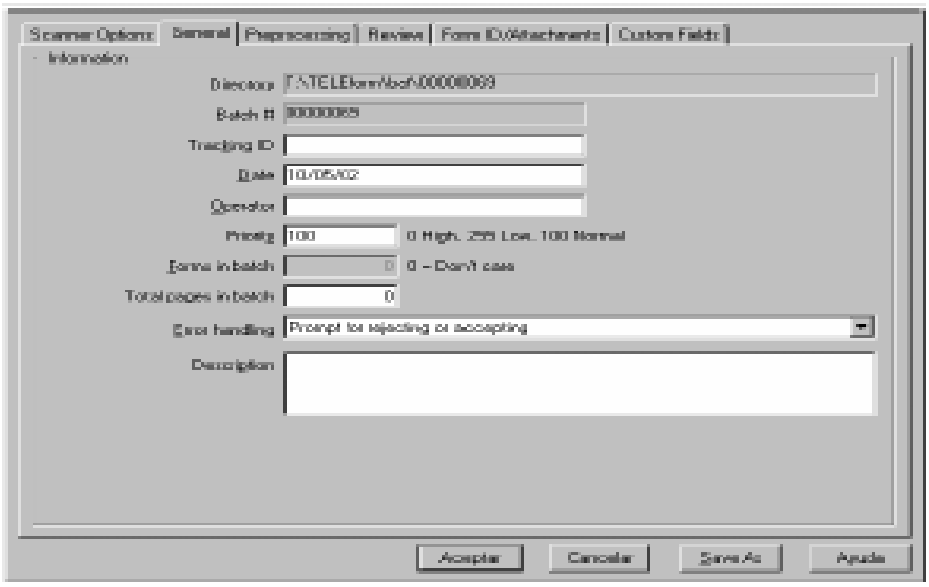


Figura 14. Estación de escaneo de TELEFORM, es desde aquí donde se administran los lotes que han sido escaneados.



- Control de proceso: lleva el control sobre el estado de los lotes procesados, rendimientos de los verificados, grado de reconocimiento y otras estadísticas vitales para el procesamiento. Se pueden extraer reportes como:
  a) Rendimiento de sistema: este reporte ayuda en el análisis de la eficiencia operativa del sistema TELEFORM
  b) Resumen de evaluaciones de las formas: provee información del tiempo que toman las formas para ser evaluadas por el lector
  c) Corrección de formas: facilita identificar en que modalidades de corrección los operadores sobresalen en productividad y en cuales necesitan más practica y entrenamiento

d) Detalle de las formas: este reporte puntualiza las ineficiencias en el diseño de la forma
e) Rendimiento de operadores: le permite identificar la eficiencia de los operadores en la verificación, así como si están corrigiendo con eficiencia o necesitan atención
f) Resumen de lotes: Estadísticas útiles si se están procesando uno o más lotes y se desea evaluar el estado del proceso.

- Lector: lo conforman los motores que utilizan los algoritmos de reconocimiento. Es el encargado de interpretar el contenido de documentos digitalizados (faxes, imágenes y datos en formularios) mediante la técnica de reconocimiento neurológico Tri-CR.

- Verificación de lectura: es él modulo con el cual el verificador puede confirmar los datos que no se han podido reconocer. Se responsabiliza de mostrar la imagen de los documentos con la variable marcada por el lector de TELEFORM a fin de que el operador pueda corregir o confirmar el dato. Utiliza técnicas que imitan en forma rítmica y sincrónica, la forma en que la mente humana ve y procesa datos.

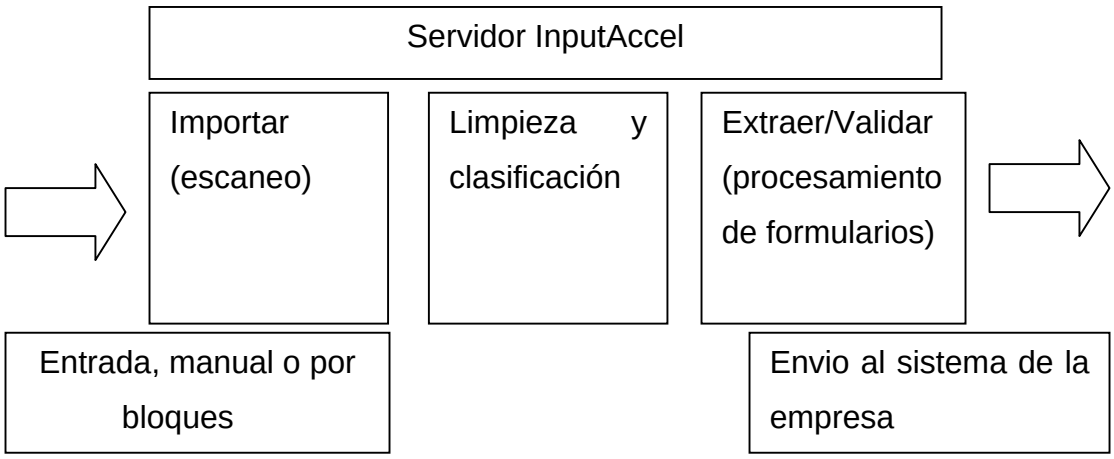
- Conectividad externa: es el componente con el cual Teleform genera los *batch* revisados y reconocidos a otras herramientas, por ejemplo: SQL, Oracle, Access, XML, ASCII, CSV y SPSS

4.1.2.3 InputAccel

Desarrollado por Captiva Software como una herramienta reconocida por su amplia flexibilidad de configuración y su arquitectura modular, siendo la principal causa de su éxito el flujo de captura (*CaptureFlows)* lo cual permite diseñar una solución especifica de acuerdo a las reglas del negocio, además de su facilidad para adaptarse a la tecnología existente y a futuras necesidades.

Gobernando los módulos, cargas de los clientes y el *captureFlows* se encuentra el servidor de InputAccel, que funciona como un administrador virtual y que automáticamente envía la información a los clientes apropiados, mientras que el servidor los supervisa, encargándose de balancear las cargas de trabajo, a fin de mejorar la productividad del personal encargado y eliminar los cuellos de botella.

Figura 15. Estructura básica de Servidor InputAccel



La arquitectura de InputAccel la conforman los siguientes módulos:

- Importar, este modulo es el encargado de obtener la información de fuentes externas, correos, documentos, etc.
- Limpieza y clasificación, que se encarga de quitar las imperfecciones y clasificar los proyectos según las reglas del negocio.
- Extraer y validar, que se encarga de obtener de las imágenes aquellas regiones que forman parte del documento definido y cuya interpretación es la que se requiere

Luego de estos módulos esta en modulo encargado de migrar la información ya sea a texto, CSV o bien a alguna herramienta de cualquier otro proveedor cuyo formato sea bien conocido

4.2 Casos de evaluación de las aplicaciones OCR

4.2.1 Procesamiento de pruebas de aptitud académica de la Universidad de San Carlos de Guatemala.

Durante los últimos y los primeros meses de cada año en la Universidad de San Carlos de Guatemala se efectúan pruebas de aptitud académica en la Unidad de Bienestar Estudiantil.

Estas pruebas tienen como objetivo el mostrar a los estudiantes cuales son las carreras son las mas adecuadas para ellos en base a factores como la habilidad numérica, y su capacidad de razonamiento verbal. Según fuentes de este departamento, el número de pruebas procesadas en la unidad durante los últimos meses del 2004 superó las 12,000, y al final del periodo de pruebas procesan más de 55,000 pruebas en promedio.

4.2.1.1 Descripcion del procedimiento actual de procesamiento de las pruebas de aptitud vocacional

Debido a la gran cantidad de pruebas y a la creciente demanda de los resultados de estas pruebas, el departamento de procesamiento de las pruebas comenzó a utilizar hace un par de años, TELEFORM 8.1, aunque hasta el momento solo una de sus computadoras posee la licencia de este software, también poseen un escáner de alta velocidad, con alimentación automática de 65 paginas por minuto, lo cual ha hecho que su capacidad para obtener los resultados de estas pruebas haya aumentado considerablemente en comparación del software que utilizaban antes llamado SCANTRON, el cual era puramente un reconocedor de marcas, y cuyo porcentaje de error era de un 20%, es decir que solamente el 80% de las pruebas eran consideradas como leídas sin presentar ningún error.

En este escenario, TELEFORM ha demostrado una gran flexibilidad en cuanto al el reconocimiento de marcas y al de caracteres, al ser mas fácilmente configurable la capacidad para definir parámetros como la intensidad, la validación de marcas y la programación de *scripts* de validación.

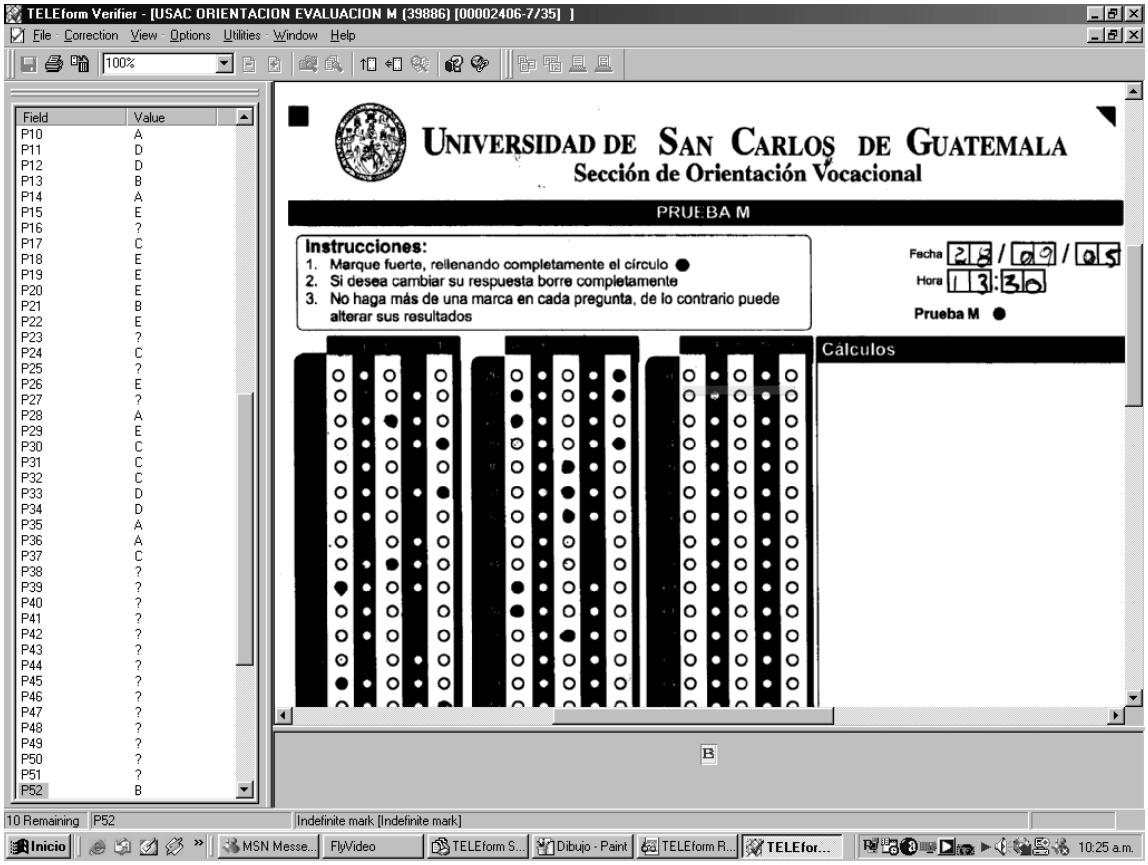
Utilizando TELEFORM es posible obtener, en el peor de los casos, un 2% de error, que representa un 98% de efectividad en la lectura de las pruebas, y si se considera que dicho porcentaje de error puede verse influenciado directamente por factores ajenos a la capacidad del motor OMR – OCR, por ejemplo, la limpieza del estudiante al llenar la prueba, los borrones, las marcas débiles o no intencionadas en lugares inadecuados.

Aunque específicamente en OCR la capacidad de este programa se ve bastante reducida, ya que en texto impreso, se vera influenciado por el tipo de letra que le será mas difícil de reconocer.

En cuanto a texto escrito manualmente, su rendimiento es aun menor, reduciendo su capacidad hasta en un 80% de efectividad, haciendo que la mayor parte del texto escrito a mano sea despreciado, puesto que llevaría un mayor tiempo de revisión el verificar y corregir lo que el programa marca como entrada inconclusas o erróneas, aquí cabe resaltar que la efectividad del OCR se ve disminuida debido a la variedad de caligrafías existentes para cada persona, y que incluso se ve influenciada por su estado de animo, personalidad, grado de fatiga, etc.

Las correcciones de las pruebas se realizan en el modulo de verificación de TELEFORM (VERIFY), aquí es posible corregir tanto, marcas como texto si el operador así lo decide.

Figura 16. la figura muestra el modulo Verify aplicado a una boleta de Orientación Vocacional,



Luego de verificado y corregida, el archivo conteniendo las respuestas de cada formulario es exportado a un archivo de texto, en el que cada carácter representa cada una de las respuestas proporcionadas por el estudiante.

Si la respuesta está en blanco o fue tomada como irreconocible aun en la verificación, el programa le colocara algún carácter especial definido que reemplace la respuesta, en este caso el carácter escogido para valores nulos o vacios '?'.

Figura 17. la figura muestra el modulo Verify solicitando se confirmen los cambios realizados

Luego para obtener los resultado de la prueba de cada estudiante, se utiliza un programa hecho en FOX para DOS versión 2.6, en el cual se grafican los resultados obtenidos por cada prueba.

Figura 18. la figura muestra el modulo Verify solicitando se validen datos escritos de forma manual en un campo que reconoce escritura tipografica.

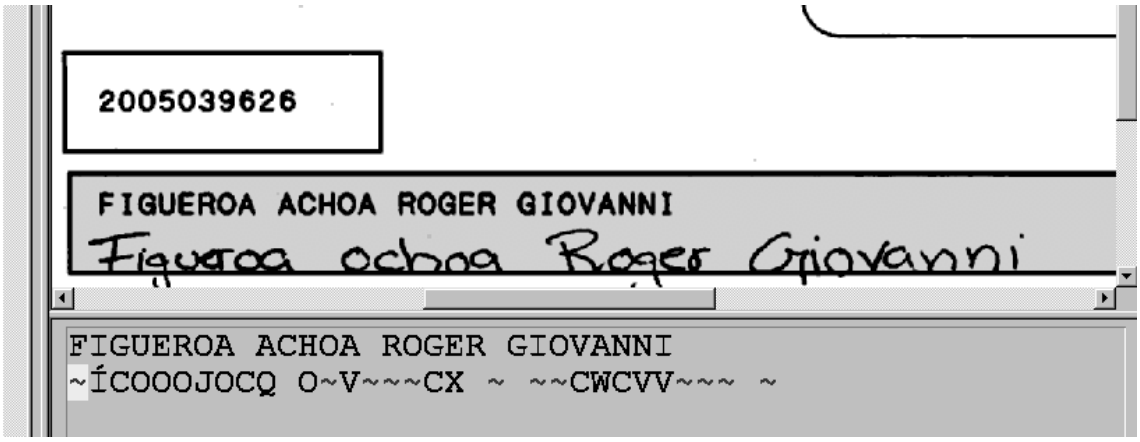


Figura 19. la figura muestra el la configuración que lleva el cuadro de texto que escanea y valida lo mostrado en la figura 18

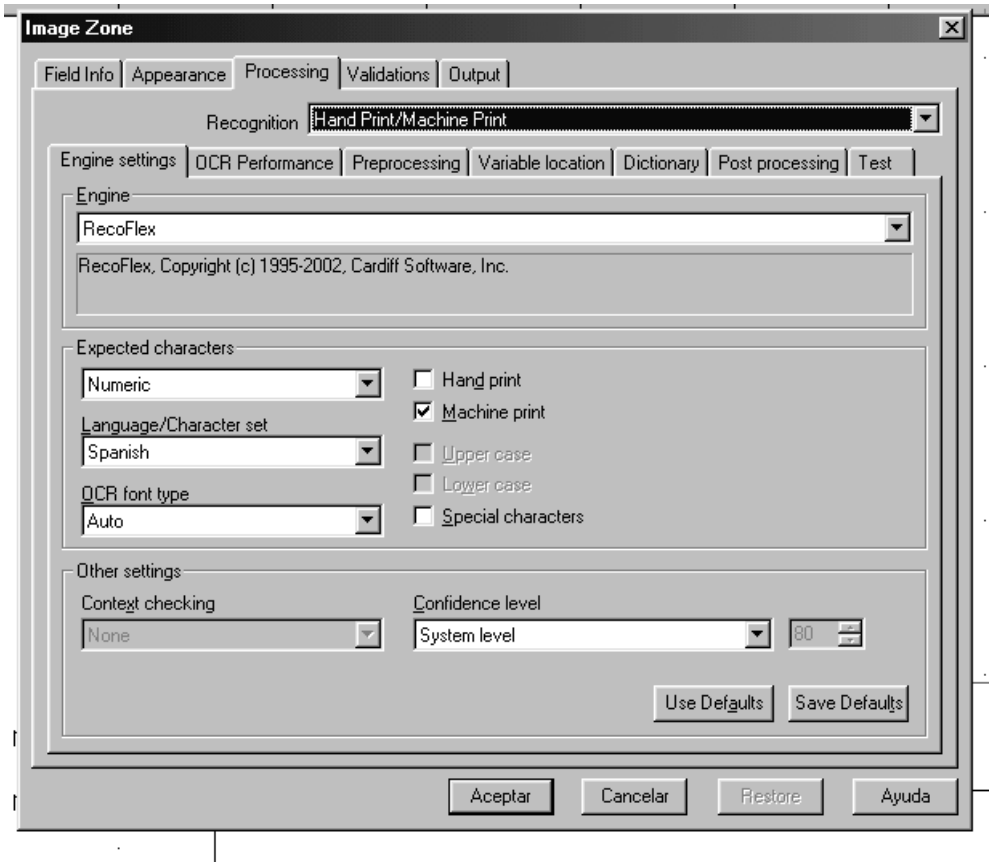
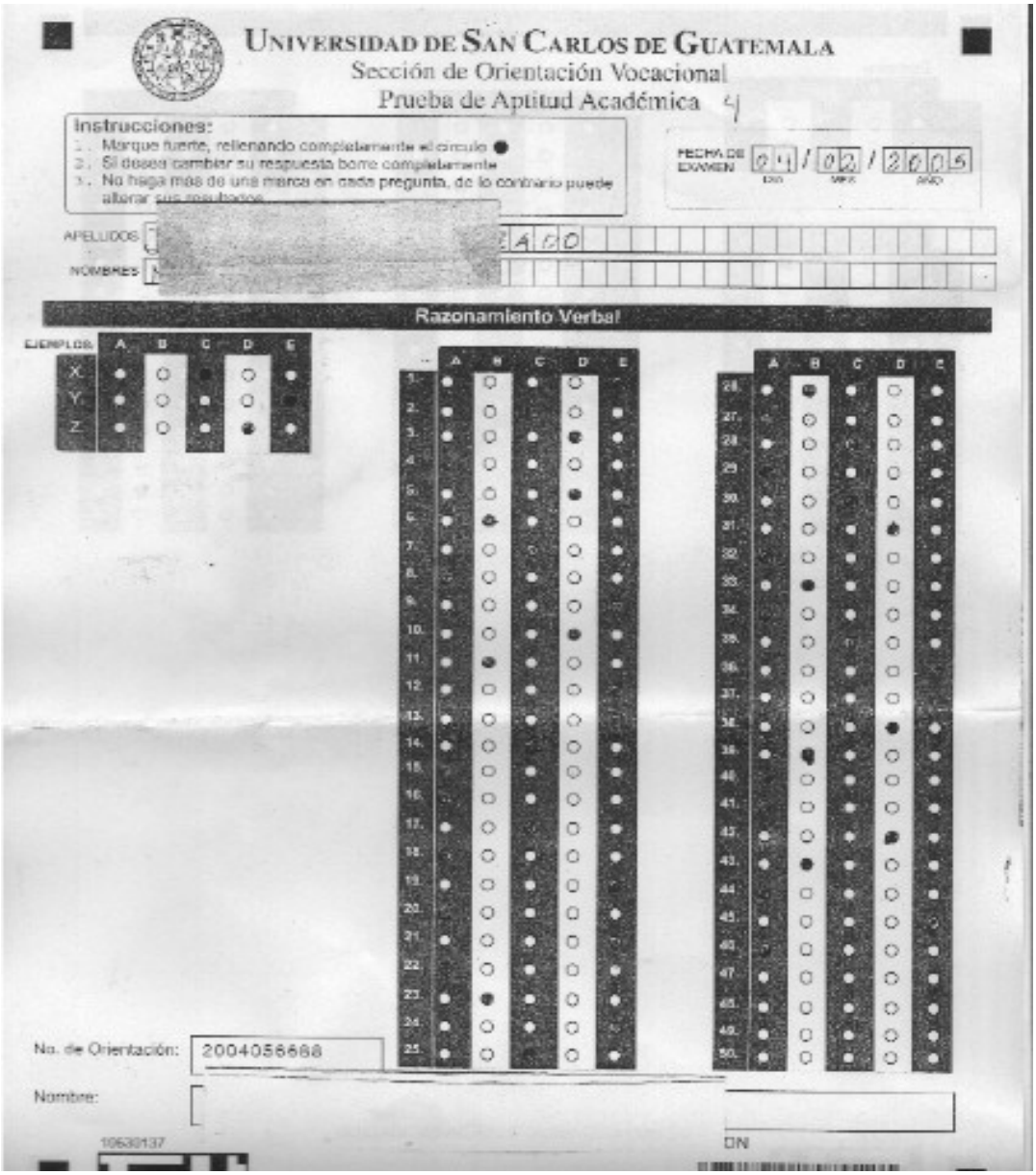


Figura 20 Prueba de aptitud academica de la sección de orientación vocacional, con respuesta en forma de marcas, que se procesan utilizando un escáner Bell-Howell 20000 FB y TELEFORM



4.2.1.2 Resultados de la evaluación de Teleform en el procesamiento de pruebas de aptitud vocacional

Para el evaluar el rendimiento de Teleform en procesar evaluaciones de aptitud vocacional se tomó un lote de 35 evaluaciones seleccionadas al azar, y que presentaban algunos de los errores mostrados en las figuras 16 a 19, las estadísticas del lote procesado se resumen en la tabla x

Tabla III.Estadísticas de producción del procesamiento de pruebas de orientación vocacional

Folletos procesados	35
Campos validados / Folleto	50
Total Imágenes digitalizadas	54
Caracteres procesados	84
Marcas reconocidas	50
Folletos procesados que necesitan revisión	10
Tiempo de escaneo por hoja (54 imágenes)	20 seg/hoja

En la tabla anterior se puede observar que el tiempo de escaneo y el numero de imágenes que se escanean es menor al que vemos en otro tipo de folletos escaneados mediante este método, esto se debe en primer lugar a que el folleto ha sido específicamente creado mediante el Teleform Designer, en segundo lugar la mayor parte de los campos que se leen del folleto o formulario son marcas (OMR) y no caracteres propiamente (OCR) las cuales son mas fácilmente reconocibles que los caracteres.

4.2.1.3 Conclusiones a las que se llegaron a partir de la experiencia de procesamiento de evaluaciones vocacionales

- El desempeño del software elegido para la evaluación probo ser bastante eficiente para el reconocimiento del formulario en mas del 98%, los factores que influenciaron este porcentaje son los siguientes:

    - El medio en el que se mantienen los formularios es mas controlado que otros en los que generalmente se tienen grandes volúmenes de documentos.
    - El casi 2% de error existente se debe a factores al momento de escanear el formulario, la mala colocacion de un formulario, borrones o marcas dobles en campos al momento de que es llenado por los estudiantes.

- El software evaluado en este caso probo ser el adecuado para obtener información de un formulario, pero cabe hacer énfasis, en que la mayoría de los campos evaluados (50 de 54) son marcas, y no caracteres.

- El tiempo que dedica el personal de procesamiento a realizar verificaciones de los lotes escaneados dura aproximadamente 1 minuto por folleto que presente algun dato no valido.

4.2.2 Censo de Piloto de Población Guatemala 2002

Actualmente algunos países se han explorando nuevas tecnologías para reemplazar el ingreso manual de información mediante operadores, a tecnologías de tipo OCR. Escanear imágenes de formularios censales ha probado ser una solución atractiva para la captura de datos, el almacenamiento de la información y la recuperación de la misma debe ser confiable, mediante métodos de control de calidad que permitan asegurar la integridad de la información capturada. En pruebas realizadas en Nueva Zelanda [6]. la tasa de exactitud de reconocimiento fue de 99.4 % para las marcas y de 74% para números. La tasa de error fue de 0.6 %, donde 0.52 % fueron marcas que no pudieron ser identificadas.

Las técnicas actuales de procesamiento para amplios volúmenes de respuestas alfabéticas, son aun difíciles de interpretar de una forma rápida y apropiada. En Guatemala, una de las primeras experiencias en el ámbito estatal en el procesamiento masivo de información tuvo lugar en el Censo Piloto de Población y Habitación se efectuó del 21 de abril al 4 de mayo de 2002. en los municipios de Santa Ana Huista y San Rafael Petzal, Huehuetenango; Jerez y El Adelanto, Jutiapa, Tamahú y Lanquín, Alta Verapaz.

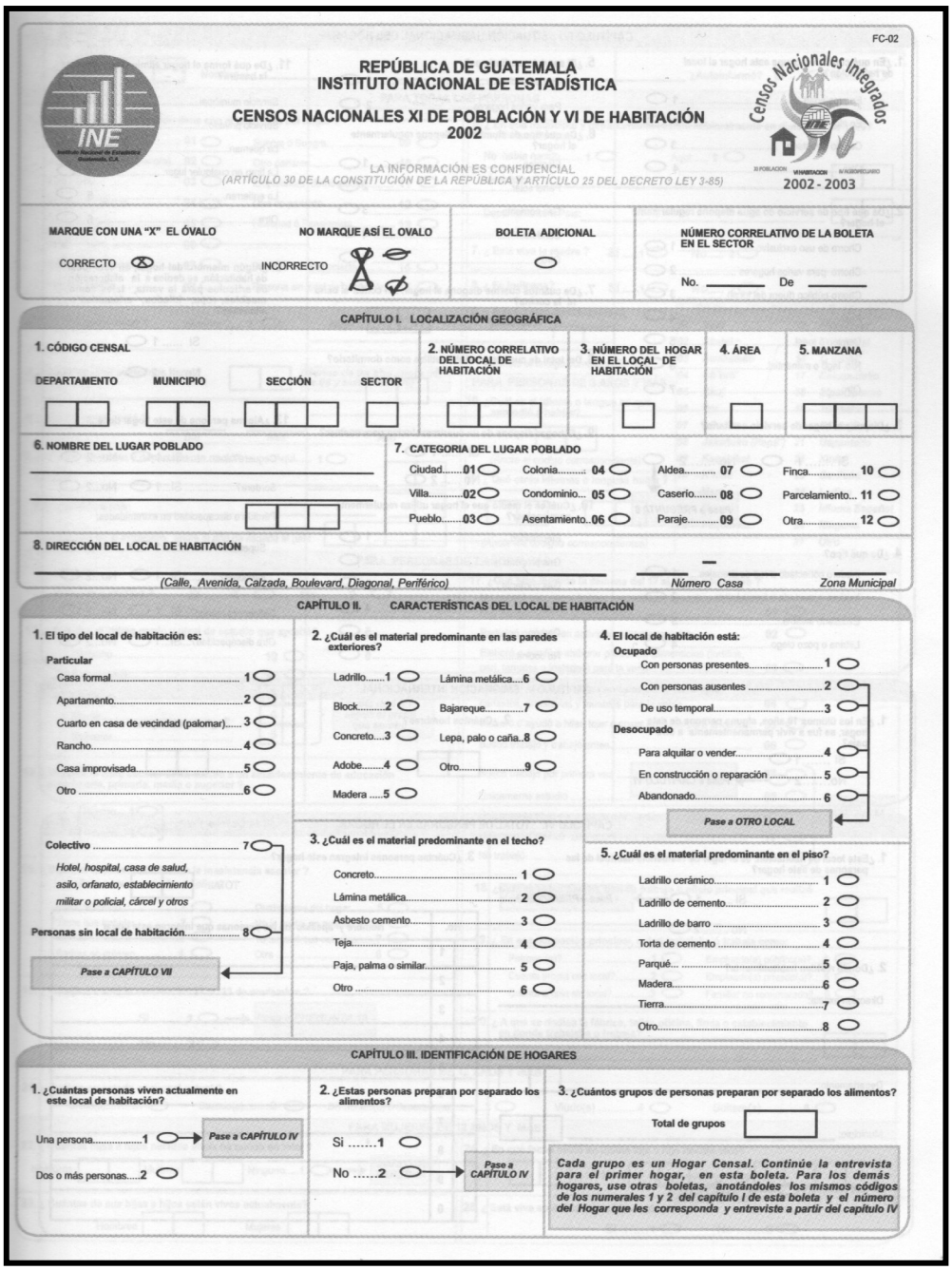
En esta oportunidad el Instituto Nacional de Estadística aparte de utilizar el método tradicional y probado de digitación, se lanzo a utilizar las nuevas tecnologías de procesamiento de datos, utilizando dispositivos ópticos de reconocimiento de caracteres.

En tal sentido, se hizo necesario diseñar un plan de procesamiento de datos con objetivos y estrategias bien definidas para normar las diferentes actividades que el proceso implica en el censo piloto. La digitalización de la información de las boletas censales tenia como objetivo, poder darle un tratamiento informático a los datos censales.

La elección de las tecnologías adecuadas, que cumplieran con los requerimientos de exactitud y rapidez en la digitalización de la boleta censal, fue el primer reto enfrentado por el departamento de procesamiento de datos. Para ello se realizó la digitalización de las boletas del Censo Piloto. Se debía probar por medio de un ejercicio real (Censo Piloto), qué tecnología cumplía la exactitud, rapidez y factibilidad presupuestaria que se requeriría para cumplir en corto tiempo la digitalización de las boletas censales, así como también, probar procesos post-digitalización como el tratamiento del dato para la generación de los resultados finales, el cual era la meta final del proyecto de Censos Nacionales.

La lectura óptica de las boletas reduciría la posibilidad de error y el tiempo de procesamiento, lo que mejoraría la calidad de la información. Esta tecnología, tiene un porcentaje de reconocimiento de 99.9% en el caso de marcas y de 98% en el de caracteres.

Figura 21.Boleta Censal, censo de Población y Habitación, Guatemala 2002, Capitulos I, II y III (Ubicación geográfica y características del local de habitación, emigración internacional)



FC-02

REPÚBLICA DE GUATEMALA
INSTITUTO NACIONAL DE ESTADÍSTICA

CENSOS NACIONALES XI DE POBLACIÓN Y VI DE HABITACIÓN
2002

LA INFORMACIÓN ES CONFIDENCIAL
(ARTÍCULO 30 DE LA CONSTITUCIÓN DE LA REPÚBLICA Y ARTÍCULO 25 DEL DECRETO LEY 3-85)

Censos Nacionales Integrados
2002 - 2003

MARQUE CON UNA "X" EL ÓVALO — CORRECTO ⊗ — INCORRECTO
NO MARQUE ASÍ EL OVALO
BOLETA ADICIONAL ○
NÚMERO CORRELATIVO DE LA BOLETA EN EL SECTOR — No. _____ De _____

CAPÍTULO I. LOCALIZACIÓN GEOGRÁFICA

1. CÓDIGO CENSAL — DEPARTAMENTO | MUNICIPIO | SECCIÓN | SECTOR
2. NÚMERO CORRELATIVO DEL LOCAL DE HABITACIÓN
3. NÚMERO DEL HOGAR EN EL LOCAL DE HABITACIÓN
4. ÁREA
5. MANZANA

6. NOMBRE DEL LUGAR POBLADO

7. CATEGORIA DEL LUGAR POBLADO

Ciudad....01	Colonia....04	Aldea....07	Finca....10
Villa....02	Condominio...05	Caserío....08	Parcelamiento...11
Pueblo....03	Asentamiento..06	Paraje....09	Otra....12

8. DIRECCIÓN DEL LOCAL DE HABITACIÓN
(Calle, Avenida, Calzada, Boulevard, Diagonal, Periférico) — Número — Casa — Zona Municipal

CAPÍTULO II. CARACTERÍSTICAS DEL LOCAL DE HABITACIÓN

1. El tipo del local de habitación es:
Particular
Casa formal....1
Apartamento....2
Cuarto en casa de vecindad (palomar)....3
Rancho....4
Casa improvisada....5
Otro....6
Colectivo....7
Hotel, hospital, casa de salud, asilo, orfanato, establecimiento militar o policial, cárcel y otros
Personas sin local de habitación....8
Pase a CAPÍTULO VII

2. ¿Cuál es el material predominante en las paredes exteriores?

Ladrillo....1	Lámina metálica....6
Block....2	Bajareque....7
Concreto....3	Lepa, palo o caña..8
Adobe....4	Otro....9
Madera 5	

3. ¿Cuál es el material predominante en el techo?
Concreto....1
Lámina metálica....2
Asbesto cemento....3
Teja....4
Paja, palma o similar....5
Otro....6

4. El local de habitación está:
Ocupado
Con personas presentes....1
Con personas ausentes....2
De uso temporal....3
Desocupado
Para alquilar o vender....4
En construcción o reparación....5
Abandonado....6
Pase a OTRO LOCAL

5. ¿Cuál es el material predominante en el piso?
Ladrillo cerámico....1
Ladrillo de cemento....2
Ladrillo de barro....3
Torta de cemento....4
Parqué....5
Madera....6
Tierra....7
Otro....8

CAPÍTULO III. IDENTIFICACIÓN DE HOGARES

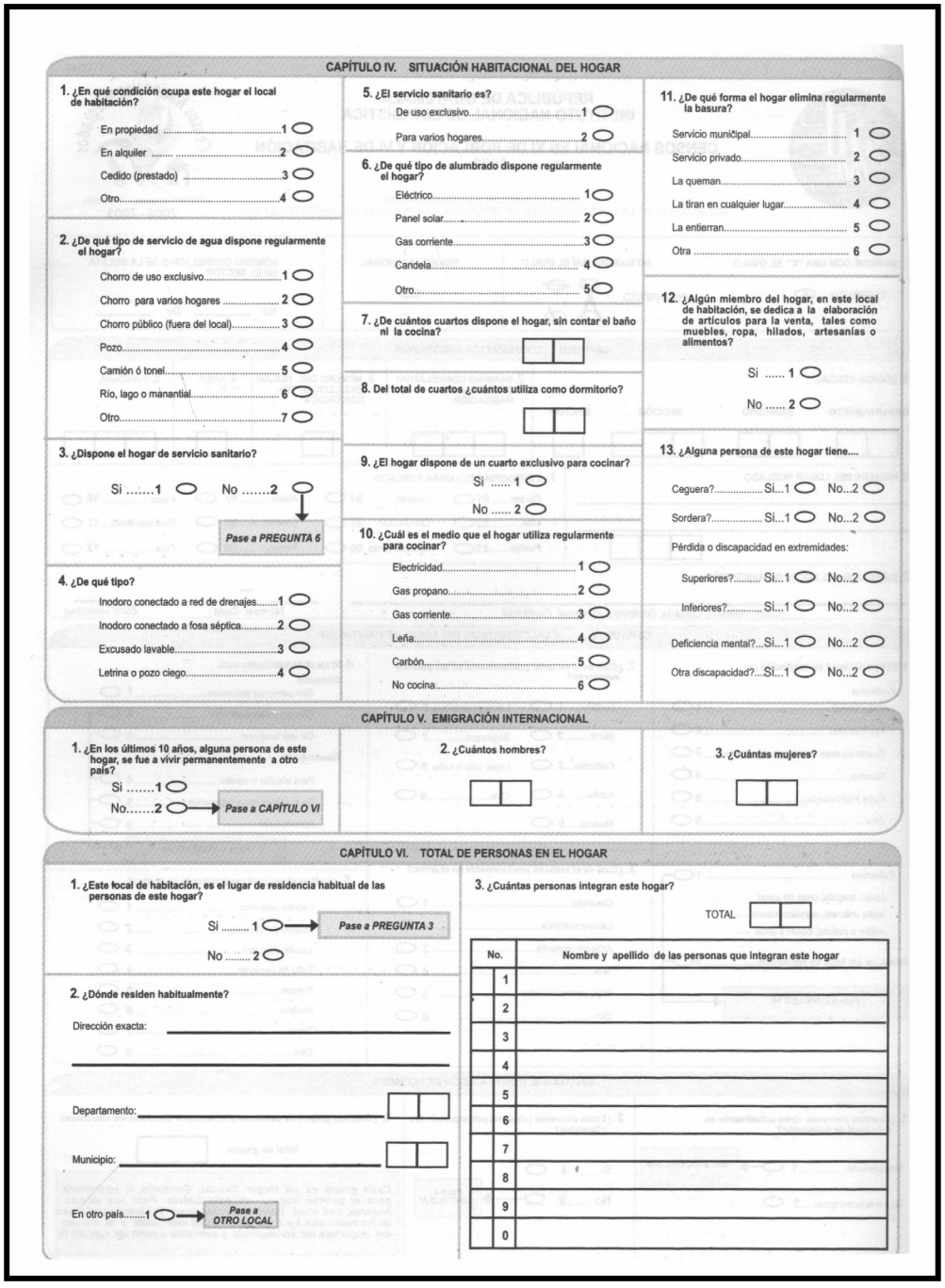
1. ¿Cuántas personas viven actualmente en este local de habitación?
Una persona....1 → Pase a CAPÍTULO IV
Dos o más personas....2

2. ¿Estas personas preparan por separado los alimentos?
Si 1
No ....2 → Pase a CAPÍTULO IV

3. ¿Cuántos grupos de personas preparan por separado los alimentos?
Total de grupos

Cada grupo es un Hogar Censal. Continúe la entrevista para el primer hogar, en esta boleta. Para los demás hogares, use otras boletas, anotándoles los mismos códigos de los numerales 1 y 2 del capítulo I de esta boleta y el número del Hogar que les corresponda y entreviste a partir del capítulo IV

Figura 22.Boleta Censal, censo de Población y Habitación, Guatemala 2002, Capitulos IV, V y VI (Situación habitacional del hogar, Emigración internacional)



CAPÍTULO IV. SITUACIÓN HABITACIONAL DEL HOGAR

1. ¿En qué condición ocupa este hogar el local de habitación?
   - En propiedad 1
   - En alquiler 2
   - Cedido (prestado) 3
   - Otro 4

2. ¿De qué tipo de servicio de agua dispone regularmente el hogar?
   - Chorro de uso exclusivo 1
   - Chorro para varios hogares 2
   - Chorro público (fuera del local) 3
   - Pozo 4
   - Camión ó tonel 5
   - Río, lago o manantial 6
   - Otro 7

3. ¿Dispone el hogar de servicio sanitario?
   - Si 1
   - No .....2 → *Pase a PREGUNTA 6*

4. ¿De qué tipo?
   - Inodoro conectado a red de drenajes 1
   - Inodoro conectado a fosa séptica 2
   - Excusado lavable 3
   - Letrina o pozo ciego 4

5. ¿El servicio sanitario es?
   - De uso exclusivo 1
   - Para varios hogares 2

6. ¿De qué tipo de alumbrado dispone regularmente el hogar?
   - Eléctrico 1
   - Panel solar 2
   - Gas corriente 3
   - Candela 4
   - Otro 5

7. ¿De cuántos cuartos dispone el hogar, sin contar el baño ni la cocina? ☐☐

8. Del total de cuartos ¿cuántos utiliza como dormitorio? ☐☐

9. ¿El hogar dispone de un cuarto exclusivo para cocinar?
   - Si 1
   - No 2

10. ¿Cuál es el medio que el hogar utiliza regularmente para cocinar?
    - Electricidad 1
    - Gas propano 2
    - Gas corriente 3
    - Leña 4
    - Carbón 5
    - No cocina 6

11. ¿De qué forma el hogar elimina regularmente la basura?
    - Servicio municipal 1
    - Servicio privado 2
    - La queman 3
    - La tiran en cualquier lugar 4
    - La entierran 5
    - Otra 6

12. ¿Algún miembro del hogar, en este local de habitación, se dedica a la elaboración de artículos para la venta, tales como muebles, ropa, hilados, artesanías o alimentos?
    - Si 1
    - No 2

13. ¿Alguna persona de este hogar tiene....
    - Ceguera? ..... Si...1 No...2
    - Sordera? ..... Si...1 No...2
    - Pérdida o discapacidad en extremidades:
    - Superiores? ..... Si...1 No...2
    - Inferiores? ..... Si...1 No...2
    - Deficiencia mental? ... Si...1 No...2
    - Otra discapacidad? ... Si...1 No...2

CAPÍTULO V. EMIGRACIÓN INTERNACIONAL

1. ¿En los últimos 10 años, alguna persona de este hogar, se fue a vivir permanentemente a otro país?
   - Si 1
   - No.....2 → *Pase a CAPÍTULO VI*

2. ¿Cuántos hombres? ☐☐

3. ¿Cuántas mujeres? ☐☐

CAPÍTULO VI. TOTAL DE PERSONAS EN EL HOGAR

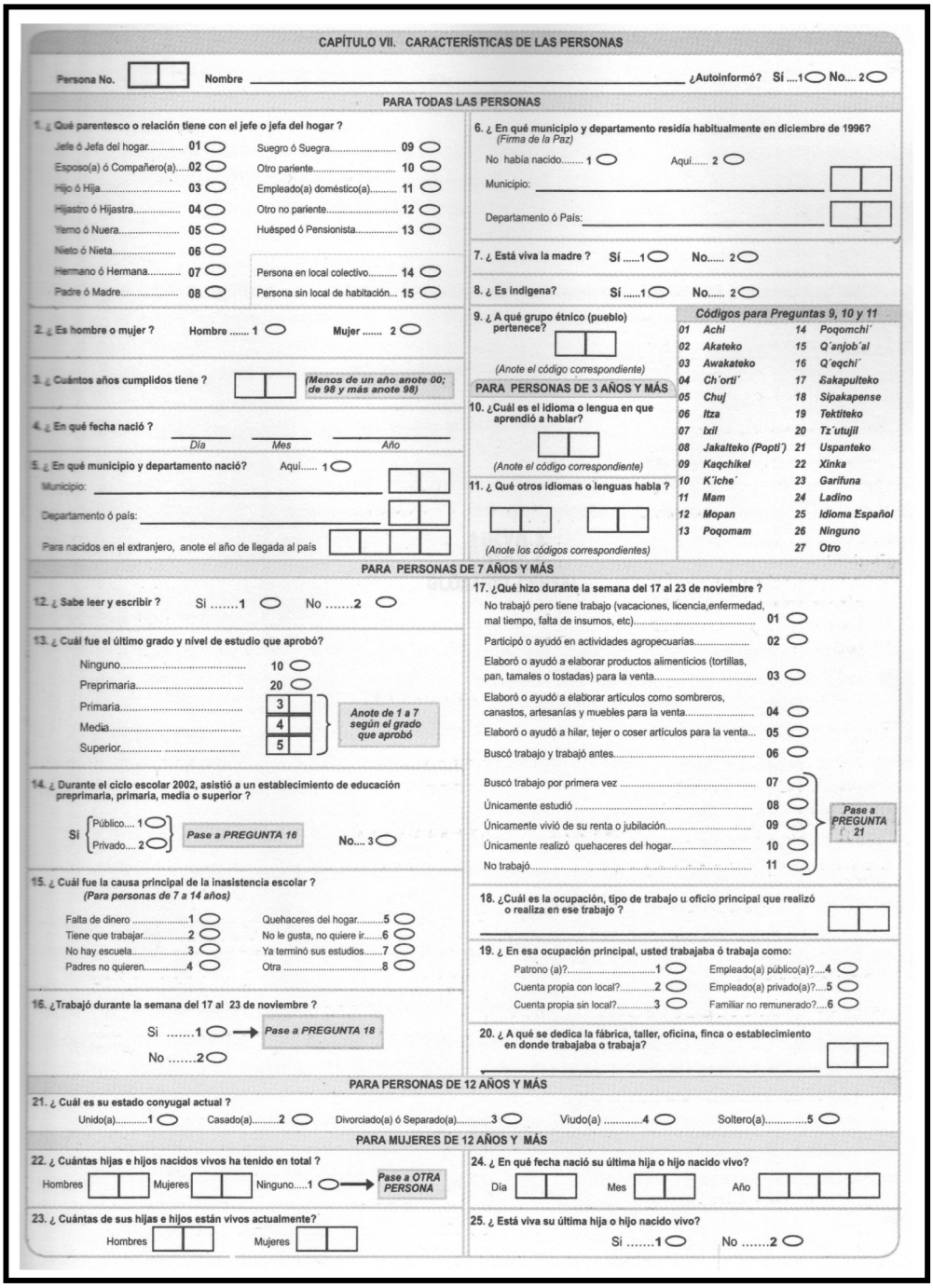
1. ¿Este local de habitación, es el lugar de residencia habitual de las personas de este hogar?
   - Si ..... 1 → *Pase a PREGUNTA 3*
   - No 2

2. ¿Dónde residen habitualmente?
   - Dirección exacta: _____
   - Departamento: _____ ☐☐
   - Municipio: _____ ☐☐
   - En otro país.....1 → *Pase a OTRO LOCAL*

3. ¿Cuántas personas integran este hogar?
   TOTAL ☐☐

No.	Nombre y apellido de las personas que integran este hogar
1	
2	
3	
4	
5	
6	
7	
8	
9	
0	

Figura 23. Boleta Censal, censo de Población y Habitación, Guatemala 2002, Capitulo VII



CAPÍTULO VII. CARACTERÍSTICAS DE LAS PERSONAS

Persona No. ☐☐ Nombre _____ ¿Autoinformó? Sí ....1 ○ No.... 2 ○

PARA TODAS LAS PERSONAS

1. ¿ Qué parentesco o relación tiene con el jefe o jefa del hogar ?

Jefe ó Jefa del hogar..... 01 ○ | Suegro ó Suegra..... 09 ○
Esposo(a) ó Compañero(a)....02 ○ | Otro pariente..... 10 ○
Hijo ó Hija..... 03 ○ | Empleado(a) doméstico(a)..... 11 ○
Hijastro ó Hijastra..... 04 ○ | Otro no pariente..... 12 ○
Yerno ó Nuera..... 05 ○ | Huésped ó Pensionista..... 13 ○
Nieto ó Nieta..... 06 ○
Hermano ó Hermana..... 07 ○ | Persona en local colectivo..... 14 ○
Padre ó Madre..... 08 ○ | Persona sin local de habitación... 15 ○

2. ¿ Es hombre o mujer ? Hombre ..... 1 ○ Mujer ..... 2 ○

3. ¿ Cuántos años cumplidos tiene ? ☐☐ *(Menos de un año anote 00; de 98 y más anote 98)*

4. ¿ En qué fecha nació ? _____ Día _____ Mes _____ Año

5. ¿ En qué municipio y departamento nació? Aquí..... 1 ○
Municipio: _____ ☐☐
Departamento ó país: _____ ☐☐
Para nacidos en el extranjero, anote el año de llegada al país ☐☐☐☐

6. ¿ En qué municipio y departamento residía habitualmente en diciembre de 1996? *(Firma de la Paz)*
No había nacido..... 1 ○ Aquí..... 2 ○
Municipio: _____ ☐☐
Departamento ó País: _____ ☐☐

7. ¿ Está viva la madre ? Sí .....1 ○ No..... 2 ○

8. ¿ Es indígena? Sí .....1 ○ No..... 2 ○

9. ¿ A qué grupo étnico (pueblo) pertenece? ☐☐ *(Anote el código correspondiente)*

PARA PERSONAS DE 3 AÑOS Y MÁS

10. ¿Cuál es el idioma o lengua en que aprendió a hablar? ☐☐ *(Anote el código correspondiente)*

11. ¿ Qué otros idiomas o lenguas habla ? ☐☐ ☐☐ *(Anote los códigos correspondientes)*

Códigos para Preguntas 9, 10 y 11

01	Achi	14	Poqomchi´
02	Akateko	15	Q´anjob´al
03	Awakateko	16	Q´eqchi´
04	Ch´orti´	17	Sakapulteko
05	Chuj	18	Sipakapense
06	Itza	19	Tektiteko
07	Ixil	20	Tz´utujil
08	Jakalteko (Popti´)	21	Uspanteko
09	Kaqchikel	22	Xinka
10	K´iche´	23	Garífuna
11	Mam	24	Ladino
12	Mopan	25	Idioma Español
13	Poqomam	26	Ninguno
		27	Otro

PARA PERSONAS DE 7 AÑOS Y MÁS

12. ¿ Sabe leer y escribir ? Si .....1 ○ No .....2 ○

13. ¿ Cuál fue el último grado y nivel de estudio que aprobó?

Ninguno..... 10 ○
Preprimaria..... 20 ○
Primaria..... ☐ 3
Media..... ☐ 4
Superior..... ☐ 5
Anote de 1 a 7 según el grado que aprobó

14. ¿ Durante el ciclo escolar 2002, asistió a un establecimiento de educación preprimaria, primaria, media o superior ?

Sí { Público.... 1 ○ / Privado.... 2 ○ } *Pase a PREGUNTA 16* No.... 3 ○

15. ¿ Cuál fue la causa principal de la inasistencia escolar ? *(Para personas de 7 a 14 años)*

Falta de dinero .....1 ○ | Quehaceres del hogar.....5 ○
Tiene que trabajar.....2 ○ | No le gusta, no quiere ir.....6 ○
No hay escuela.....3 ○ | Ya terminó sus estudios.....7 ○
Padres no quieren.....4 ○ | Otra .....8 ○

16. ¿Trabajó durante la semana del 17 al 23 de noviembre ?
Si .....1 ○ → *Pase a PREGUNTA 18*
No .....2 ○

17. ¿Qué hizo durante la semana del 17 al 23 de noviembre ?

No trabajó pero tiene trabajo (vacaciones, licencia,enfermedad, mal tiempo, falta de insumos, etc)..... 01 ○
Participó o ayudó en actividades agropecuarias..... 02 ○
Elaboró o ayudó a elaborar productos alimenticios (tortillas, pan, tamales o tostadas) para la venta..... 03 ○
Elaboró o ayudó a elaborar artículos como sombreros, canastos, artesanías y muebles para la venta..... 04 ○
Elaboró o ayudó a hilar, tejer o coser artículos para la venta... 05 ○
Buscó trabajo y trabajó antes..... 06 ○
Buscó trabajo por primera vez ..... 07 ○
Únicamente estudió ..... 08 ○
Únicamente vivió de su renta o jubilación..... 09 ○
Únicamente realizó quehaceres del hogar..... 10 ○
No trabajó..... 11 ○
Pase a PREGUNTA 21

18. ¿Cuál es la ocupación, tipo de trabajo u oficio principal que realizó o realiza en ese trabajo ? ☐☐

19. ¿ En esa ocupación principal, usted trabajaba ó trabaja como:

Patrono (a)?.....1 ○ | Empleado(a) público(a)?....4 ○
Cuenta propia con local?.....2 ○ | Empleado(a) privado(a)?....5 ○
Cuenta propia sin local?.....3 ○ | Familiar no remunerado?....6 ○

20. ¿ A qué se dedica la fábrica, taller, oficina, finca o establecimiento en donde trabajaba o trabaja? ☐☐

PARA PERSONAS DE 12 AÑOS Y MÁS

21. ¿ Cuál es su estado conyugal actual ?
Unido(a).....1 ○ Casado(a).....2 ○ Divorciado(a) ó Separado(a).....3 ○ Viudo(a) .....4 ○ Soltero(a).....5 ○

PARA MUJERES DE 12 AÑOS Y MÁS

22. ¿ Cuántas hijas e hijos nacidos vivos ha tenido en total ?
Hombres ☐☐ Mujeres ☐☐ Ninguno.....1 ○ → *Pase a OTRA PERSONA*

23. ¿ Cuántas de sus hijas e hijos están vivos actualmente?
Hombres ☐☐ Mujeres ☐☐

24. ¿ En qué fecha nació su última hija o hijo nacido vivo?
Día ☐☐ Mes ☐☐ Año ☐☐☐☐

25. ¿ Está viva su última hija o hijo nacido vivo?
Si .....1 ○ No .....2 ○

La boleta censal estaba compuesta en su gran mayoría de campos numéricos y marcas, que representaban la información (codificada, en algunos casos) de las viviendas, locales de habitación y personas, siendo los campos alfa numéricos pocos, pero debido a la importancia de la información que estos contenían se hizo necesario que el sistema OCR / ICR cumpliese con las características que se mencionan antes para un sistema de este tipo (reconocimiento de marcas,etc.)

Tabla IV.Comparaciones de rendimiento al procesar imágenes.

	ICR-OCR	**OMR**
Reconocimiento de escritura a mano	Si-ICR	No
Reconocimiento de escritura a maquina	Si-OCR	No
Reconocimiento de marcas tipo burbuja	Si	Si
Reconocimiento de marcas tipo “X”	Si	No
Reconocimiento de códigos de barra	Si	Si
Requiere marcas de tiempo en formulario y ID´s de formas	No	Si
Requiere marcas de registro	Si	No
Exactitud	Arriba de 98% con altibajos	Consistente en 99.9%
Almacenamiento y recuperación electrónica	Si	No
Ingreso por FAX	**Si**	No
Velocidad	Sin limite, dependiendo de la configuración de la Red y cantidad de Scanner	1,500 – 10,000 Hr/Scanner

4.2.2.1 Procesamiento de Boletas Censales utilizando Teleform.

El implementación del procesamiento óptico de las boletas que se detalla a continuación es la presentada por la empresa Moore que participo en el proceso de licitación para el Censo Nacional procesando para el Censo Piloto un lote de 4,500 viviendas aproximadamente en el municipio de Lanquín, Alta Verapaz, esta empresa presento su solución utilizando Teleform,

Todo el proceso de digitalización hasta entrega de datos a Procesamiento de Datos de cada uno de los proveedores, fue meticulosamente examinado por observadores quienes formaban parte del personal de Censos Nacionales Integrados, los que debían documentar toda la metodología que los proveedores debían seguir para lograr un archivo magnético de los datos de las boletas censales.

Con el fin de unificar criterios de supervisión, se elaboró un listado de puntos a evaluar y preguntas a responder sobre los diferentes aspectos de la tecnología. Para cada uno de los puntos se elaboró un cuadro de control para las anotaciones y observaciones necesarias y su posterior revisión y calificación.

El procedimiento de procesamiento de datos del censo piloto utilizado fue el siguiente:

1. Procesamiento y recuperación de imágenes de parte del proveedor
2. Selección aleatoria de dos (2) sectores procesados por el proveedor
3. Digitación tradicional de los sectores seleccionados
4. Verificación de los sectores digitados
5. Ejecución de programa que realiza el pareo de datos por variable (sectores procesados vrs. Sectores digitados)
6. Listar diferencias encontradas en pareo de datos
7. Verificar diferencias contra boleta física
8. Realizar correcciones de digitación (si hubiese) y regresar al paso no. 5 hasta que no existan errores de digitación.
9. Presentar resultados finales del pareo de datos

Figura 24. Diagrama de flujo para el censo piloto (Procesamiento Óptico)

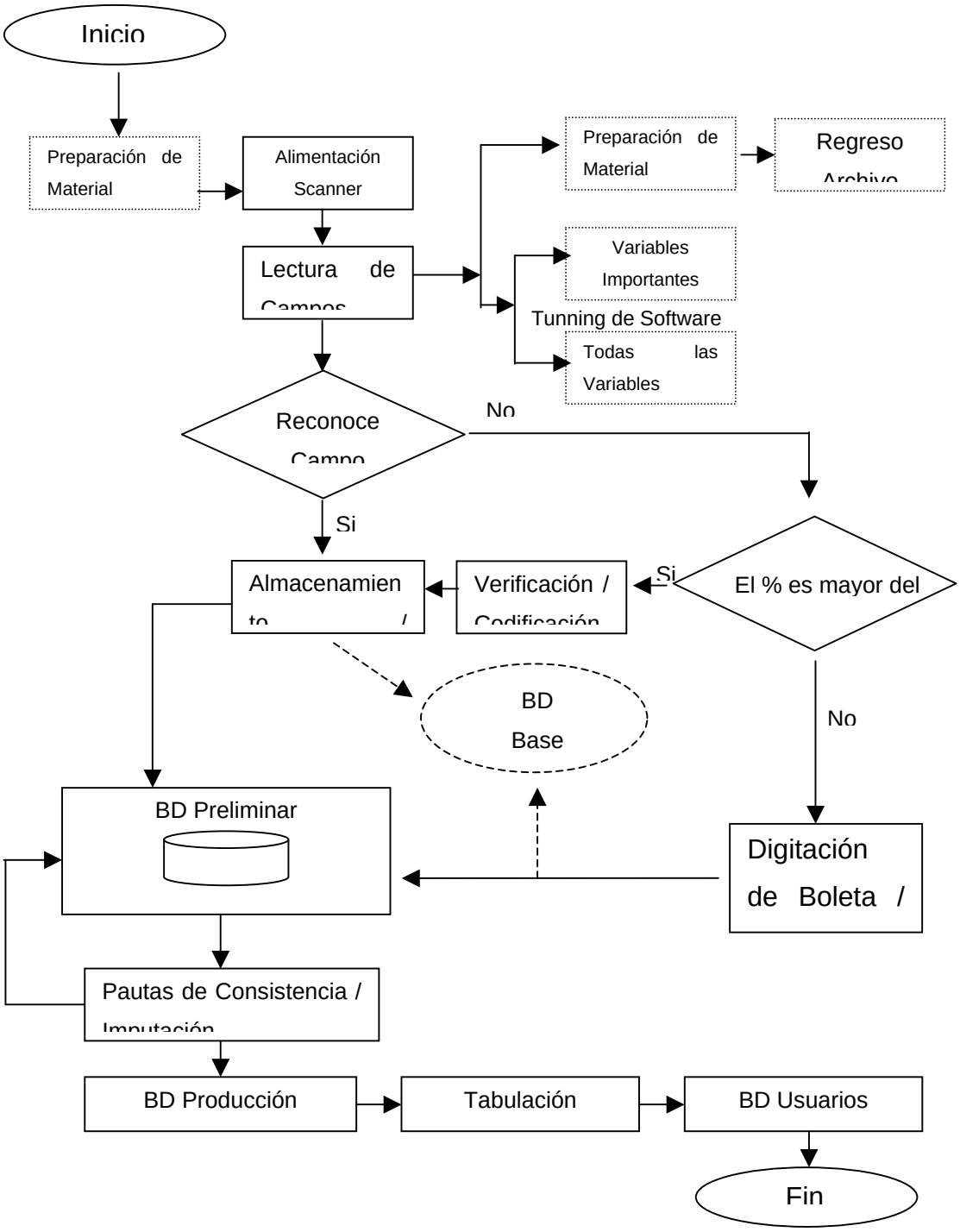


Tabla 5.Estadísticas de procesamiento en el Censo Piloto utilizando Teleform

	Piloto 1	Piloto 2
Total folletos procesados	3,109	3,109
Campos validados / Folleto	444	532
Total Imágenes digitalizadas	18,654	18,654
Total caracteres procesados	2,001,311	2,104,745
Total lotes procesados	90	90

Tabla 6.Estadísticas de producción en el Censo Piloto utilizando Teleform

	Piloto 1	Piloto 2
Inicio Fin	13/05 12:45 PM 14/05 3:00 PM	15/05, 4:00 PM 16/05 4:45 PM
Preparación de documentos	TET = 9hrs	
Digitalización	2.82 (110 Imag/Min)	
Reconocimiento (Teleform Reader)	8.2 Seg / Imag	
Verificación (Horas por operador) a 10,000 PPH	52:57 Horas 400 horas	77:26 Horas 421 Hrs
Verificación - Imag/Hora Max Min Promedio	 630 330 414	 336 180 228

4.2.2.2 Conclusiones a las que se llegaron a partir de la experiencia del Censo Piloto.

- La tecnología de procesamiento óptico de imágenes, probo ser funcional, pero se presentaron problemas que dificultaron tanto el escaneo como la interpretación de las imágenes escaneadas, como:
    - Suciedad en las boletas censales, este factor puede dificultar tanto el levantado de la información por falta de legibilidad como por manchas que pudieron ser interpretadas como marcas.
    - Boletas maltratadas, arrugadas o dobladas lo cual no permite que las boletas resbalen por los *scanners* y que estas sean operadas.
    - La calidad de la escritura de los empadronadores resulta ser un factor importante, puesto que los variados estilos caligráficos llevo a que algunos tipos de escritura fueran de difícil interpretación, unidos a los factores antes mencionados.
    - Los lotes (bolsas) de boletas censales estaban incompletos o con hojas adicionales que requerían un proceso adicional de verificación, o bien la digitación de toda la boleta, aumentando el tiempo de procesamiento
- La tecnología de procesamiento óptico de imágenes, es exacta, por lo que puede ser utilizada en proyectos que como este, tiene grandes magnitudes de boletas a procesar en poco tiempo.
- La tecnología de procesamiento óptico de imágenes, es una tecnología de alto costo en su ejecución.
- La digitación tradicional es menor en costo y contribuiría a la generación de empleo por la intensiva mano de obra.

4.3 Consideraciones finales al procesar información mediante el reconocimiento óptico de caracteres

Debido al volumen de información que se necesita procesar y el costo de toda la operación, es necesario considerar el tener algún tipo de respaldo en caso de que el equipo falle. Los *scanners* utilizan enormes cantidad de espacio en disco, para poder almacenar imágenes. Un cuestionario censal esta compuesto de entre 2 y 4 hojas. Para contener la información disponible en ambos lados de la hoja, son necesarios cerca de 80 Kbytes, aunque esto también dependerá de la resolución que la imagen tenga y que necesite(n) el (los) motor(es) necesarios para reconocer esta imagen.

Si bien el uso de técnicas de compresión de disco son útiles, también ha de considerarse el tiempo que le tomará al motor el descomprimir la información antes de poder dar una respuesta, otra posibilidad incluye el guardar únicamente las imágenes del texto que contenga la información a capturar, pero esta solución incrementa el tiempo de procesamiento. Es necesario además considerar el diseño del cuestionario como una posible fuente de errores al procesar la imagen, debe evitarse que los campos de un lado de la hoja estén en el mismo nivel que los del otro lado, ya que esto podría hacer que la sombra de alguno de los campos sea interpretada por el *scanner* como parte de la imagen que escanea.

Al utilizar *scanners* es importante el considerar que aquellas respuestas que forman parte de un campo alfa numérico no son tan fáciles de interpretar como lo serían marcas, las variables de actividad económica, profesión u ocupaciones, muestran cuan complicado puede ser este asunto.

En Latinoamérica, y más específicamente en Colombia este probó ser un problema de grandes dimensiones, inicialmente se desarrollo un procedimiento que identificara las letras que componían el campo, luego se utilizaba un diccionario para identificarla y asignarle un código, si no la encontraba este generaba un nuevo código, ahora bien si estas tenían raíces similares, esto podría llevar a un error de codificación.

Para solucionar este problema, expertos en el área de procesamiento de censos proponen, el utilizar una método híbrido de procesamiento, ya que el reconocimiento de campos alfa numéricos, no es aun tan exacto como se quisiera, un censo bien podría procesarse en dos fases, la primera todos aquellos campos numéricos o de marca, que estén dentro del cuestionario.

En una segunda fase se procesarían todos aquellos campos que estén formados por cadenas alfa numéricas de gran tamaño y completitud. Siendo esta probablemente la mejor solución para acelerar el ingreso de la información. La metodología del procesamiento de datos mediante dispositivos de reconocimiento óptico, sigue evolucionando, y los algoritmos empleados cada vez se diversifican más, optimizando las búsquedas y los tiempos de respuestas, haciendo de esta una tecnología que con el pasar de los años se volverá mas y más eficiente, haciendo que los métodos tradicionales se vuelvan obsoletos.

Si bien muchos de los factores descritos anteriormente deben ser considerados también es necesario hacer énfasis en lo que en determinado momento puede limitar el uso de estas tecnologías en el procesamiento masivo de información, este factor es el alto costo de una operación de este tipo.

Puesto que para realizar este tipo de procesamiento se debe contar con *scanners* de alta definición y velocidad que permitan digitalizar las imágenes, operadores para dicho equipo (digitalizadores), y personal de digitación que se encargue de revisar y verificar lo escaneado.

Como se puede observar en la tabla 5, tiempos de entre 52 y 77 horas para un lote de 3100 documentos aproximadamente pueden representar altos costos en concepto de pago de honorarios al personal de digitación (actualmente se cotiza la plaza temporal de digitador entre Q.3,000.00 y Q.3,360.00), el cual deberá ser seleccionado de tal forma que las verificaciones que haga, las haga en base a un criterio definido, encareciendo aun mas su mano de obra, visto de cierta forma a un grupo de 3 digitadores como el que laboro en el Censo Piloto con la solución antes descrita, les tomo aproximadamente 7 días de trabajo el verificar los 3109 documentos , a esto habría que sumarle el tiempo que lleva el escaneado y el tiempo empleado por el operador en darle mantenimiento al escáner

Si bien como se menciono anteriormente esta tecnología ha probado ser de gran utilidad para reducir tiempos, el costo a nivel del personal que la utilice puede en algún momento ser la mayor y quizás más difícil barrera que permita desarrollar este tipo de soluciones en ambientes cuyo contexto pueda reducir la capacidad de estas tecnologías.

CONCLUSIONES

La creatividad humana sigue siendo la fuente de las maravillas tecnológicas que nos rodean, si bien la tecnología descrita anteriormente, aún presenta pequeños agujeros en cuanto a su exactitud, es hoy por hoy, uno de los métodos más rápidos para el procesamiento de información, y es en donde la creatividad humana entrará en juego para hacer que el rendimiento de ésta se incremente, al mejorar los actuales algoritmos en que se basa.

En cuanto al reconocimiento de caracteres escritos a mano, cada avance que la Inteligencia Artificial haga, permitirá que, tarde o temprano, una computadora sea capaz de reconocer cuán variado puede ser el significado semántico de una frase, esto a su vez repercutirá, en las actuales restricciones que la tecnología de reconocimiento óptico tiene en cuanto a las palabras escritas.

Si bien los resultados del caso práctico no fueron los esperados, es necesario recordar que esta tecnología es hasta cierto punto nueva, al menos a nivel nacional y esto no debe ser un factor que lleve a su desuso, por el contrario, es necesario poner en práctica el uso de esta tecnología, puesto que las posibilidades de incrementar la productividad, reducir los tiempos y costos financieros, hacen que su uso sea bastante atractivo en cualquier área, social, médica, educacional, etc., que involucre el procesamiento de largos volúmenes de información.

RECOMENDACIONES

- En los cursos del pensum de la carrera de Ingeniería en Ciencias y Sistemas, es de gran utilidad en la experiencia académica del estudiante, involucrarlo en proyectos que permitan conocer y desarrollar el área de reconocimiento óptico.

- Adquirir, al igual que universidades en el extranjero que se dedican a esta área, equipo y formar grupos de investigación que permitan mejorar los actuales algoritmos empleados en la interpretación obtenida mediante el uso de tecnología OCR.

- Automatizar procesos que en Guatemala aún se hacen de forma manual, en los cuales la cantidad de información hace que se vuelvan lentos y tediosos, por ejemplo, el proceso de recepción y envío de correos, encuestas, digitación de documentos completos (libros, revistas, etc.) mediante el uso de esta tecnología.

BIBLIOGRAFÍA

- Ashlock, Dan. Data Crawlers for Simple Optical Character Recognition, Mathematics and Complex Adaptive Systems, Iowa State University. EEUU.

- Barney Smith, Elisa H. Document Scanning Defect Analysis using Bilevel Image Features, Electrical Engineering Department, Boise State University. EEUU.

- Brown, Eric W. Character Recognition by Feature Point Extraction EEUU.

- Debashish Niyogi, Sargur N. Srihari. The use of document structure analysis to retrieve information from documents in digital libraries, Center of Excellence for Document Analysis and Recognition (CEDAR). EEUU.

- Egecioglu Omer, Maximilian Ibel. Parallel Algorithms for Fast Computation of Normalized Edit Distances (Extended Abstract) Department of Computer Science University of California Santa Barbara, EEUU.

- Ellis, Carlos. New Techniques Used for Massive Data Entry.

- Hall, Cyrus. Hybrid Single-Link Clustering for Optical Character Recognition Dept. of Computer Science University of Colorado, Boulder. EEUU.

- Hammerstrom, Dan. The OGI Cognitive Architecture Project, Electrical and Computer Engineering Department, Oregon Graduate Institute, EEUU.
- Kam Ho, Tin. George Nagy, OCR with No Shape Training. Bell Labs, Dept of ECSE, Rensselaer Polytechnic Institute, EEUU.

- Kanungo Tapas, Robert M. Haralick An Automated Closed-Loop Methodology for Generating Character Groundtruth, Center for Automation Research, University of Maryland. EEUU.

- Kia, Ohmid E,.David Doermann, Rama Chellapa. Compressed Domain document retrieval and analysis, Document Processing Group, Center for Automation Research, University of Maryland. EEUU.

- Lecoq, J.C., L.Najman, O.Gibot2 and E.Trupin, Benchmarking Commercial OCR Engines for Technical Drawings Indexing, Francia

- Pérez Cortés, Juna C, Amengual Joaquim Arlandis Stochastic Error-Correcting Parsing for OCR Post-processing. Instituto Tecnológico de Informática (ITI) Universidad Politécnica, Camino de Vera , Valencia, España.

- Tong, Xiang. David A. Evans. A Statistical Approach to Automatic OCR Error Correction in Context, Laboratory for Computational Linguistics, Carnegie Mellon University Pittsburgh, EEUU.

- Ward Church, Kenneth. Murray Hill, Patrick Hanks , Word Association Norms, Mutual Information, and Lexicography Bell Laboratories, CoLlins Publishers, Glasgow, Scotland