論 文

論文69—12—4—2

한글 文字의 電子計算組織에 適應하기 위한 特徵抽出에 관한 硏究(I)

(A Method for the Recognition of Printed Korean Characters)

李 柱 根*
(Lee, Joo Keun)

要 約

우리 文字는 子母로 組合된 言語文字이기 때문에 그 數가 방대하여 數千個의 識別기구를 必要로 할뿐만 아니라 構造가 복잡하며 大部分이 類似文字이기 때문에 Pattern 認識問題에 있어서는 허다한 難點이 있다. 따라서 이들 構造上에서 오는 問題點을 分析, 評價하여 最適條件을 決定하고, 特徵抽出에 論理識別函數의 적용은 다른 文字에서는 볼 수 없는 한글 文字에 關한 限 特異한 長點으로 나타난다는 것을 確認하였다. 이 特異點을 System의 設計에 最大限으로 적용하여 3分之1 以上의 System縮小를 보았다.

認識方法으로서는 標本Pattern을 抽出해 내서 Register에 記憶한 다음 認識Matrix에 의하여 識別하였다. 識別된 文字는 判定論理에 의하여 特徵Parameter를 抽出 하였다. 理論的인 立證을 위한 몇가지의 實驗的인 檢討를 加하였으며 이過程에서 얻어진 모든 資料들은 이 分野의 硏究에 매우 有益한 기초資料를 제공할 것으로 보며, 한글文字의 Pattern認識에 관한 실마리가 잡혀졌다고 보겠다.

ABSTRACT

This paper attempts to analyize structure of the Letters for the purpose of making recognition of Han-Geul printed and described the method of recognition and design of the optimum system.

For the reason of the Consistency of Han-Geul (Korean Letters) combined with Consonants and vouels, the number of the words used in the daily living is about 2,000 words. For this reason the composition of the recognition system is complicated, and therfore, this paper is pursued to research to handle the separate way in each form of Letter between consonant and vouel. and the further description of this paper also indicates us the many parts of savings of elements when the character is extracted as logic system in Letter composition.

I. 序 論

이 硏究는 한글文字의 Pattern 認識過程에서 組織上의 問題點에 대한 評價와 特徵抽出에 대해서 檢訂하였다.

원래 文字의 自動認識에 대한 관심은 情報의 解讀, 自動Data處理과정에서 힘입게 되어 Perceptron[1] 以來 Computer에의 적용에 注目을 모으게 되었으며, Alphabet文字나 數字에 대해서는 이미 實用化하고 있다. 그러나 現段階로서는 어떤 文字이건 統一된 方式으로서 識別할 수 있는 方法은 아직 發見되어 있지 않다. 그러므로 한글文字에 대한 우리 스스로의 硏究開發을 必要로 하게 된다. 그런데 한글文字에 대한 이 分野의 硏究는 아직 되어 있지 않을 뿐만 아니라 文字組織의 認識過程에서의 問題點조차 검토

* 仁荷工大 電氣工學科
Dept of Electrical, of
Inha Institute
Technology

되어 있지 않다.

한글文字의 組織成分을 관찰하면 지금까지 알려진 모든 文字의 認識에서 사용된 文字組織과는 本質的으로 다를뿐 아니라 言語文字이기 때문에 그 數가 방대하며, 獨特한 구조와 그에 따른 問題點으로 困해서 그들 System에 直接 적용될 수 없다.

현재까지 많은 文獻에서 提案된 文字의 Pattern 認識方式으로서는 크게 區分하여 다음 두가지의 形式으로 集約된다.

(1) 文字의 統計的性質을 알고, 統計的인 方法에 의하여 認識하는 方式과 (2) 주어진 文字로부터 幾何學的인 特徵을 抽出해 내서 決定論的인 認識方法등으로 大別된다. 前者는 單能的인데 反해서 後者의 경우는 어떤 特徵을 抽出하는 것이 認識하기에 가장 有效한가는 아직 알려져 있지 않은 채로 相當히 융통성이 있기 때문에 많은 文獻 [2] [5]에서는 이에 근거를 둔 여러가지의 문제가 다루어져 있다.

Highleyman [2] 은 特徵 Image에 重點을 두고, 線形識別函數를 筆記體數字의 認識에 적용하였고, L.F Fourer [3]은 抵抗 Matrix의 認識 System에 類似文字의 識別을 위해서 Inhibitory weight를 導入하였고, Image가 Photo recepter를 넘어서 연속 相關運動을 하여 어는 순간에 형성되는 特性이 認識에 어떻게 작용하는가 하는 문제를 다루었다. J. Nilsson [4] 는 Learning machines의 理論體系를 集約하였다. 또 Udagawa [5]등은 特徵 Parameter의 線分을 抽出하고, Bayes의 決定規則을 적용하여 統計的 2段學習기구에 의하여 認識하였으며, Sakai [9]등은 文字의 特徵을 端點, 折曲點, 分岐點, Loop, 獨立點등의 5種으로 設定하고, 이들 特徵이 어떤 順序로 出現하는가를 보고 文字로 識別하는 方法을 論하였다.

本 硏究에서는 우선 한글文字의 組織成分에서 오는 認識過程에서의 問題點을 分析, 評價하여 最適條件을 決定하고, 活字體의 子母文字로부터 Mesh에 의한 標本 Pattern을 만든 다음, 그로부터 몇個의 特徵을 抽出해 내서 標本 Pattern과 비교하여 認識하는 方法을 검토하였다.

一般的으로 文字의 特徵은 標本에 따라 달라지기 쉬우며 더욱 筆記體인 경우는 사람의 習性에 따라 文字의 形態가 여러가지의 모양으로 表現되어서 文字의 變形, 크기의 不同, 位置의 變動등을 일으킬 수 있다. 또 印刷의 濃淡, 紙面에서의 back ground noise, 人工的 System Noise등 여러가지의 條件이 수반된다. 이러한 문제는 光學的인 檢出에 있어서는 현저한 영향을 주게 된다. 故로 本紙에서는 一次的으로이들의 制限條件을 억제하기 위하여 活字體의 文字를 標準으로 設定하였고, 識別不可能한 文字에 대해서는 몇가지의 規則을 적용하였다.

Ⅱ. 認識過程에서의 文字組織의 評價

한글 文字의 Pattern 認識에 있어서 文字構造의 特異性에서 오는 問題點을 검토하기 위하여 우선 文字의 組織成分을 分析, 評價하여 本硏究의 基本設定을 하였다. 이는 對象文字의 基準設定 및 理論구성의 基本要困과 가장 有力한 識別函數의 Factor를 찾아내기 위한 때문이다.

모아 쓰기와 풀어 쓰기 表現의 組織分類는 Teletypwriter등에서 사용되고 있지만 그것은 다만 기계적인 分類法에 지나지 않고, Pattern 認識에 입각한 것은 아니다. 따라서 여기서 論議되는 것은 主로 認識관점에서 본 組織上의 문제점 만을 다룬다.

1. 모아 쓰기의 경우 ; 子母의 文字를 각각 한개씩만으로서 構成되는 모아쓰기의 文字에 대해서 관찰하면 同一한 母音이 文字마다 반복되어 出現한다. 즉

가=ㄱ⊕ㅏ 거=ㄱ⊕ㅓ
나=ㄴ⊕ㅏ 너=ㄴ⊕ㅓ
다=ㄷ⊕ㅏ 더=ㄷ⊕ㅓ
⋮ ⋮
하=ㅎ⊕ㅏ 허=ㅎ⊕ㅓ
고=ㄱ⊕ㅗ 기=ㄱ⊕ㅣ
노=ㄴ⊕ㅗ 니=ㄴ⊕ㅣ
도=ㄷ⊕ㅗ 디=ㄷ⊕ㅣ
⋮ ⋮
호=ㅎ⊕ㅗ…히=ㅎ⊕ㅣ

표—1

등과 같이 文字의 한劃씩만 다를뿐 特性의 90%

以上이 完全히 同一形으로 나타난다. 즉 가→카 나→다, 다→마, 사→자, 자→차,……등과 같이 한 文字의 特徵을 다른 文字가 거의 內包하고 있기 때문에 認識 System으로서 識別하기에 가장 不利한 條件을 주며, 誤判別의 原因이 된다. 뿐만 아니라 이들 類似文字가 한두個만이라면 그 部分에 대해서만 特別고려를 하면 識別이 가능하겠지만 大部分이 類似文字로 나타나기 때문에 앞에서 論議된 線形識別函數의 적용은 매우 어렵게 된다.

以上은 우리 文字의 구조에서 나타나는 Pattern 認識에 있어서의 큰 缺點의 하나이다.

2. 두째로 표—1에서 보인바와 같이 母音이 각 文字에서 10個씩 반복하여 14個組를 형성한다. 이는 子母를 獨立的 文字로 하고, 認識기구를 구성할 때와 비교한다면 無味한 素子가 所要되어 素子數가 20倍에 이른다는 直視的인 結論이 나온다. 이는 기구의 복잡성을 일으킬뿐만 아니라 無味한 중복을 초래하여 科學的이 아니다. 따라서 經濟的인 面과 誤判讀을 減少시키기 위하여서는 子母文字를 獨立的으로 表現하는 方式이 識別에 훨씬 效果的이란 것을 暗示하여 준다.

3. 받침을 고려한 모아쓰기의 경우

이 경우는 더욱 복잡하여 지며, 각종 問題點을 안고있다. 子音과 母音의 組合文字의 最小數는 2要素로 區分되고, 最大數는 5要素로 區分된다. (6요소 경우도 고려되지만 일반적으로는 흔히 쓰이지 않는다) 즉 子母 각 한個씩만으로서 組合된 2要素 區分형 A (가, 나, 다, ……히, 고, 노, 구, 뉴……호,)와 쌍받침을 고려한 5要素區分형 B (땡, 빽, 쌩, …)의 경우를 생각할 수 있다. 이外에도 3要素區分형 (공, 달, 해, …등)과 4要素區分형 (빵, 딱, 닭, 책, …등)의 여러형이 存在하지만 위에서 이미 모아쓰기가 적합하지 않다는 結論이 내려졌으므로 最小와 最大區分형 이外의 組織에 대해서는 本硏究밖에의 것으로써 다만 System 구성의 評價를 위해서 最小 2要素區分형과 最大 5要素區分형에 대해서만 관심의 대상이다. 그것은 한 文字를 識別하기 위한 認識 System의 구성은 最大要素의 文字를 기준으로 해야하기 때문에 最小 2要素형도 5要素형으로 해야 하므로 방대한 素子數를 要한다.

따라서 풀어쓰기 24字로서 識別할 때에는 24個의 識別기구이면 足하지만 모아쓰기에서는 System全體가 近 400倍에 이르고 있다. 이는 認識기구로서 가장 不利한 두번째의 큰 缺點이다.

以上에서 모아쓰기가 認識過程에서 가장 不利함을 알게되고, 子母를 獨立的으로 表現하는 풀어쓰기 文字를 標本으로 삼는것이 가장 有效하고도 간결한 方法이란 근거가 서게된다. 반드시 모아써야만 된다는 理論的 근거는 發見 할 수 없으며, 科學的으로 機械化의 最短距離는 풀어쓰기 表現法이다. 풀어쓰니까 읽기 힘든다 하는것은 習慣에 의한 것 뿐이다.

4. 풀어쓰기의 경우

풀어쓰기가 決定的으로 有利할 것이란 추론은 이미 섰지만 子母만의 경우도 여러 問題點이 있다. 한글의 子母文字는 單純하면서 變化없는 似類性으로 困하여 識別函數의 判別에 허다한 難點이 있다. 實用化에 있어서는 數字와 몇個의 記號를 並用해야 할것이므로 數字 및 記號와도 同一形이 또 나타난다. 線形識別函數를 적용하기 극히 어렵고, 順序理論函數의 적용은 그림 1 (a), (b)와 같이 非對稱의 文字에서는 각 文字의 部分 Pattern Ω_0 $\Omega_1\Omega_2\Omega_3$의 出現 順序가 獨立的이기 때문에 非同期論理구성이 용이하지만 (c) (d)등과 같은 對稱文字에서는 $\Omega_1\Omega_2\Omega_3$을 相互獨立的으로 할수는 없다. $\Omega_1=\Omega_3$가 되어 버리기 때문에 이들의 內部狀態의 遷移圖를 그려보면 Ω_3와 Ω_1의 位置를 교환해야 한다. 또 그림

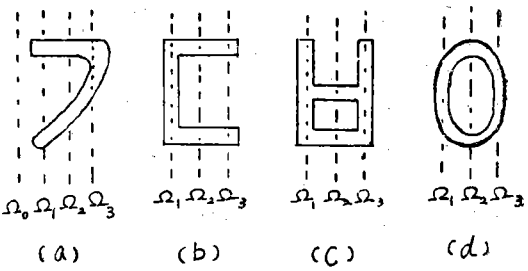


그림 1. 對稱, 非對稱文字
Symmetrical and Unsymmetrical Letters.

(b)에서 部分 Pattern $\Omega_1, \Omega_2, \Omega_3$의 出現順序가 記號「 」에서의 出現順序와 같기 때문에 두個以上의 判定이 동시에 나타나서 識別할 수 없게 된다. 또 數字 1,0과 文字 1,0에서도 마찬가지 이다. 이와 같이 識別이 不可能한 文字에 대해서는 다음에 따로 定한 規則에 의하여 識別할 수 있게 된다.

5. 長 點

ㄴㄷ, ㄹㅌ, ㅁㅇ, ㅈㅊ, …등과 같이 유사성 때문에 識別에 어려워 지는 短點도 되지만 文字가 大部分이 縱橫의 두線分으로서 組合되어 있기 때문에 이들을 列를 Scanning 즉 等分했을 때 各 區分點에서의 部分 Pattern $\Omega_{i(j)}$ (i=1.2.3…, j=1.2.3…N)가 同一函數로 抽出된다는 것이 極히 注目된다. 즉 ㄴ, ㄷ, ㅌ 등에서 $\Omega_2=\Omega_3=\Omega_4\cdots\cdots=\Omega_n$, ㄹ에서는 $\Omega_2=\Omega_3=\Omega_4\cdots=\Omega_{n-1}$, ㅁ, ㅂ에서는 $\Omega_1=\Omega_n$, $\Omega_2=\Omega_3=\cdots=\Omega_{-1}$등과 같이 한 文字에서 多數의 同一한 識別函數를 抽出해낼 수 있다는 것은 이를 論理函數에 적용하여 論理構成理論을 導入한다면 장치의 大幅的인 縮少가 可能하여 진다는 것을 뜻하게 된다. 이는 한글 文字에 關한 限 큰 長點으로 등장한다는 것을 보이고 있으며, 本 硏究에서 變化없는 한글子母 文字의 短點을 逆用하므로서 長點으로 轉換시킨 着想인 것이다. 다음 設計例에서 立證된다.

Ⅲ. 認識方式

그림 2은 한글文字의 特徵을 抽出해내기 위한

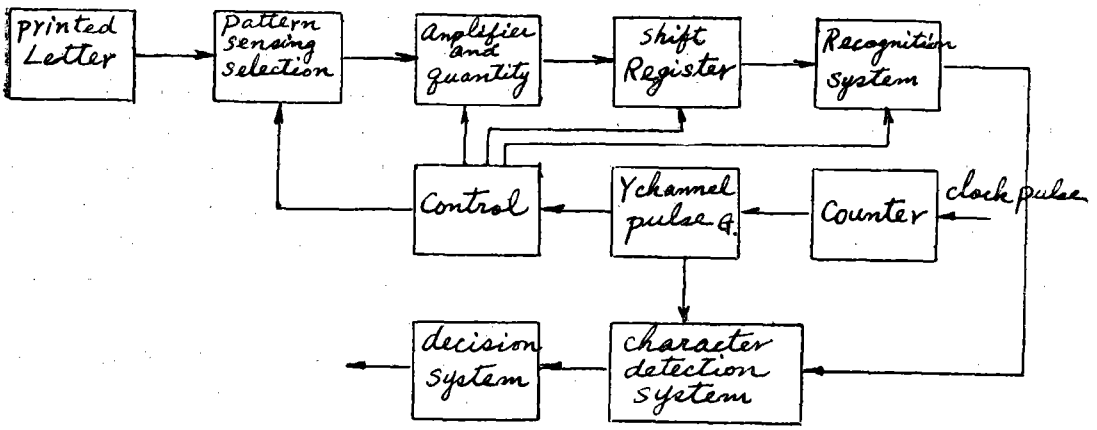


그림 2. 文字의 認識기구의 機能 Block diagram.
Blacck Schematic diagram of the recognition System

方式의 機能을 表示한 block diagram이다. 이 System의 동작원리는 (1) 우선 주어진 文字를 관측하고. 感覺 System에 의하여 文字 Pattern을 抽出하여 電氣的인 信號로 變換한다. (2) 이信號를 增幅, 量子化한 다음 이를 Shift Register에 順次的으로 옮겨서 一時記憶하여 둔다. (3) 다음 記憶된 文字 Pattern을 認識 System에서 解讀하여 未知의 文字를 識別한다. (4) 識別된 文字를 Y channel 선택 Pulse $y_1, y_2, y_3 \cdots y_7$로서 特徵 patten을 선택하여 正規位置에 分配한 다음 判定回路로서 特徵 Parameter를 抽出해 낸다. (5) 또 抽出端에 Coding하면 電子計算機에 直結 할 수 있고, 必要에 따라서는 記憶장치에 記憶해 둘 수도 있고, 천孔할 수도 있다. (6) 特別히 마련된 display에 연결하면 原文字가 表現된다.

이 System의 入力文字의 圖形은 縱橫m×n로 區分된 3×5mesh에서 각 mesh에 걸리는 文字의 黑部分이 50%以上일 때는 그 mesh는 黑으로 하고, 그 以下일 때는 白으로 하여 그 mesh 値는 1과 0으로 對應시킨다.

그림 3은 주어진 活字體에서 3×5 mesh로 量子化하여 2値 Code로서 表現한 文字의 一次特徵系의 標本 Pattern이다. 이 標本 Pattern의 生

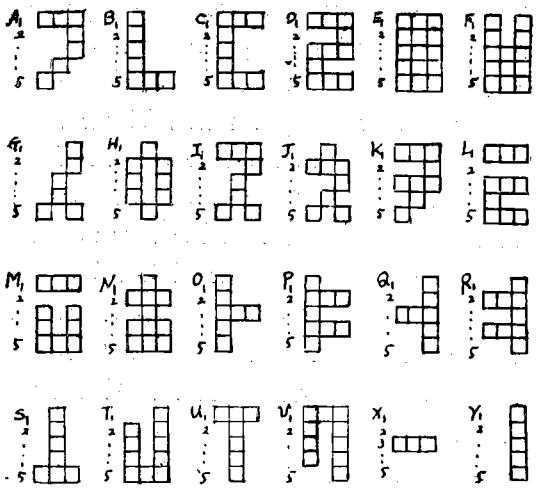


그림 3. 標本 Pattern Sampled Patterns

起順序로서 Register에 記憶 시켰다가 Binary Sequence filter에서 飜譯되서 文字가 識別된다.

文字 Pattern의 發生方法으로서는 높은 分解能을 얻을 수 있는 光電方式을 적용하였다. 그림 4와 같이 文字紙를 기계적인 구동기구에 의하여 光電變換素子群 $X=(X_1, X_2, X_3 X_n)$의 Head밑으로 走査시키고, 强烈한 光源에 의하여 文字上에 投射시켜 反射 image를 一列로 羅列한 mosaic cell에서 檢出한다. 文字紙가 移出함에

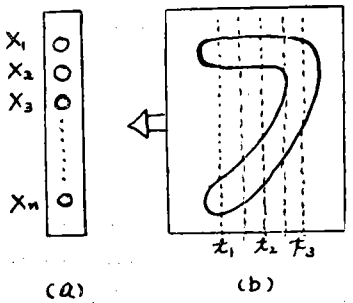


그림 4. Pattern 檢出方法

따라 Channel $X_1, X_2, X_3 \cdots$, X_n에서 檢出된 文字 image는 時間的으로 變하는 電氣的 信號로 變換된다.

그런데 文字圖形은 2次元平面上의 黑白의 點의 集合으로 고려되며, 2次元平面上의 圖形은 元來 時間과는 하등의 관계를 갖고 있지 않다. 따라서 文字圖形의 信號에 대해서 信號의 開始, 계속 時間등 時間的으로 제한을 주는 同時式보다는 信號에 weight를 주는 Analog式이 훨씬 有理하다. 그런데 2値論理함數를 적용할 때에는 時間개념이 導入되어야 한다. 더욱 $X_1, X_2, X_3 \cdots X_n$의 時時刻刻의 信號列를 全部檢出해서 順序回路에 加한다는 것은 回路를 극도로 복잡화 시킨다. 그래서 檢出된 pattern 信號는 最初의 黑點이 檢出되었을 때 t_1 時刻에서 Reading cycle이 開始되도록 하고, 一定한 時間間隔(t_1, t_2, t_3)에서 各 cell로부터의 信號를 增幅, Sampling시킨다. 또 같은 文字에 대해서도 印刷된 位置, 印刷條件등에 따라 Noise를 수반하여 모든 信號는 變質되므로 도리어 한 文字에서 많은 情報量을 抽出하기 보다는 몇個의 特徵點만을 抽出해서 識別하는 편이 간결하고도 效果的인 方法이다. 이러한 개념은 信號檢出의 情報理論의 過程에도 볼 수 있다.

우리의 日常경험에서 볼 때 文字의 一部가 떨어져 나갔거나, 찌그러져서도 우리는 文字를 識別할 수 있다.

이것은 경험적인 관측도 되겠지만 그보다도 源泉的인 要因은 그의 特徵이 남아있기 때문인 것으로 본다. 아마 사람의 얼굴에서 特徵點만 뽑아 버린다면 모든 얼굴은 같을지도 모른다. 이러한 기본적인 개념의 導入은 비록 기계적인 認識기구일지라도 特徵抽出이 有效하리라는 論證이 간다.

따라서 그림 4에서와 같이 주어진 文字에서 3個의 特徵的인 斷面만을 抽出하기로 한다. 一列로 配列한 垂直 cell群 $X_1, X_2, \cdots\cdots, X_n$을 3回의 Sampling $(t_1, t_2. t_3)$을 하여 抽出되는 信號를 增幅, 量子化한다. 이 量子量은 黑白 1,0으로 對應되며, 黑白의檢出을 위한 光學的인 感覺回路는 그림 5에 圖示하였다. 그림 (a)에서 光 Image가 入射되면 T_r의 出力에 Pulse가 나타나는데, 左側의 T_r은 Sampling用으로서 時刻 t_1, t_2, t_3에서 Sampling pulse가 加해지면 檢出信號가 出力端에 나타난다.

그림 (b)는 Pattern $X=(X_1, X_2, \cdots\cdots, X_n)$의 受光部를 한個의 T_r로서 Sampling하여 量子化된 並列 Pattern $J_1, J_2, \cdots\cdots, J_n$를 檢出해 내는 方法이다. 즉 並列로 檢出된 量子化된 Pattern을

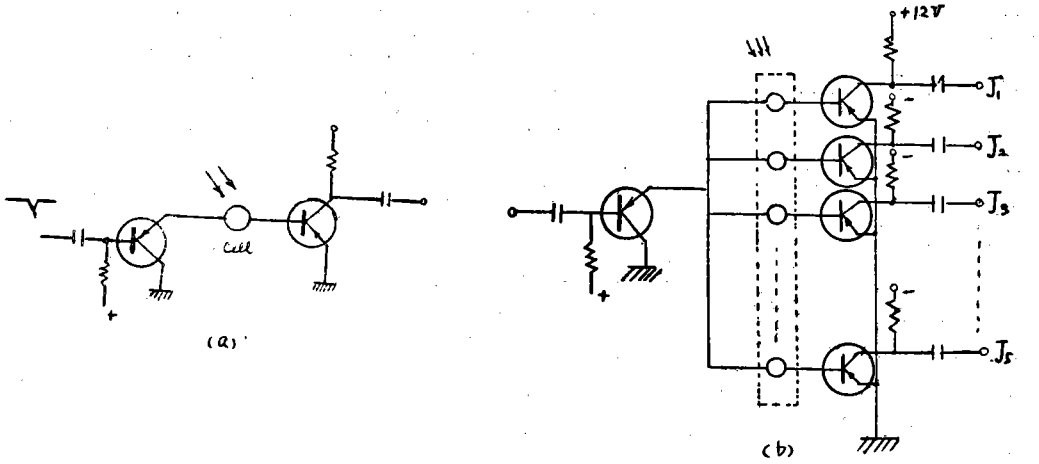


그림 5. Pattern Detection circuits

直並列로 된 Shift Register에 Sampling time에 對應되는 列에 Shift 시켜서 一時記憶하여 둔다.

다음은 記憶해둔 Pattern을 D.T.L matrix*로서 구성되어진 認識 System에서 識別하여 文字를 번역한다. 즉 Register에 記憶된 未知의 文字 Pattern이 認識 System에서 標準 Pattern配列과 一致하는 出力線에만 應答하도록 設計하고 餘他의 出力線에서는 應答하지 않도록 한다. 물론 이때에 다른 文字의 pattern이 加해지면 이 出力線은 應答하지 않도록 한다.

만약 未知의 文字 Pattern (ㄱ를 표현하는 Pattern 配列)이 Register에서 認識 System에 옮겨졌다면 이에 對應되는 認識 System의 ㄱ의 應答線에만 電壓이 發生되고 다른 모든 應答線은 Short 상태를 이루어 出力은 零으로 된다. 따라서 入力文字는 ㄱ라는 것이 判定된다. 또 ㄴ, ㄷ, ㄹ, ……, 등의 各 未知의 文字 pattern配列이 順次的으로 認識 System의 入力에 加해졌을 때에도 해당 應答端 以外의 端子에서는 零電壓이 나타나도록 設計한다.

以下 모든 文字에 대해서도 같은 方法이 적용된다. 그런데 앞에서 지적된 部分 Pattern의 識別函數가 같은 順序로 나타나서 두個 以上의 判定이 동시에 이루어져서 識別이 不可能한 文字에 대해서는 다음과같은 規則을 적용한다.

1) A, B 두 文字를 가정하여 A의 Pattern을 黑으로 했을 때에는 B는 白으로 한다.

2) 반대로 A 文字의 Pattern을 白으로 했을 때는 B는 黑으로 한다. 이와 같은 規則을 적용하면 두 유사 文字의 識別이 可能하여 진다. 그런데 本節에서 識別된 임이의 文字 f(r)은 Register에서 記憶된 文字 Pattern과 matching되어서 文字를 解讀하고는 있지만 外見上으로는 ㄱ이거나 ㄴ이거나 어느 文字에서나 同一한 階段電壓으로 解讀하고는 있다. 그래서 外見上 아무런뜻도 없이 보이는 階段電壓에 지나지 않지만 이들 電壓은 확실히 入力文字의 에 一致된 應答端에만 나타난 信號이 Pattern므로 그들은 해당 文字의 情報量을 의미하고 있음이 틀림없다. 그러나 이들 階段電壓을 다시 原文字의 Pattern으로 再現시키는 2次的인 조작을 해야 비로소 原 Pattern을 表現하는 特徵 Pattern이 再現된다. 이러한 조작은 다음 몇가지 過程을 거침으로써 이루어 진다.

Ⅵ. 特徵(Pattern)의 抽出

Pattern 認識問題에 있어서 文字가 內危하고

* 設計圖는 Appendix A—4(Ⅰ)(Ⅱ)에 주어진다

있는 모든 情報量을 全部抽出해서 識別하는 것은 非能率的이며 부차적인 問題를 일으키는 要因이 된다는 것은 앞에서 지적한바 있다. 그래서 그림 3의 1次特徵系의 標本 Pattern속에서 몇個의 대표적인 特徵點만을 선정해서 文字의 特徵 Pattern을 抽出해 내면 그림 6과 같은 文字의 特徵이 抽出된다. 이는 標本에 3×5mesh에서 抽出해낸 子母 24字에 대한 特徵 Pattern으로서 우리 文字의 子母 24字에 대한 特徵을 全部包含하고 있다. 여기서 注目할 點은 24個의 文字中에서 7個의 特徵 Pattern의 部分集合이 抽出되었는데, 이들 中에서 $\omega_1, \omega_2, \omega_3, \omega_4$의 5個의 Parameter는 文字 ㄱ은 標本 Pattern에 대한 特徵 Parameter이고, 나머지 ω_5, ω_6, ω_7이 23個의 文字에 대한 特徵 Pattern을 代表하고 있다. 故로 이 경우 文字 ㄱ는 大部分의 文字에 대한 特徵을 가지고 있다는 事實이 밝혀졌다. 이와 유사한 文字 ㅈ, ㅊ도 같은 情報量을 가지고 있음을 알 수 있다. 또 ㄱ이외의 23個文字에서 3個의 特徵만이 抽出되었으므로 이 23個의 文字는 相互共通된 特徵을 거이 갖고 있다는 것을 뜻한다. 이와 같은 現象은 우리 文字가 구조상으로 變化가 적은 극히 單調롭다는 것을 나타내며, 大部分의 文字가 縱과 橫의 線分만으로서 한 文字에 線分 한個씩만 增加시켜 다른 文字가 형성되는 多樣性이 없는 組織으로 되어 있기 때문인 것으로 본다. 때문에 前述한 認識에서의 缺함으로 나타난다. 結局 24個文字에서 7個의 部分 Pattern으로 完全히 表現되는데, 그림 6은 한글子母 文字의 最小의 特徵 Parameter이다. 이 以下가 되면 文字를 再現시킬 수 없다. (다만 ω_6, ω_7은 ω_5와 같게 할 수는 있다). 再現된 文字美를 살리기 위해서는 5×7mesh로 하는것이 좋으나, 材料를 最小로 하기 위해서 3×5mesh로한 것이다.

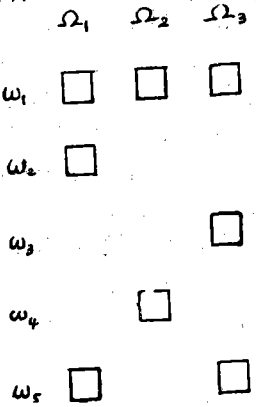


그림 6. 文字의 特徵Fig 6. Letter charac-ters.

다음 前記의 認識 System에서 識別한 文字 f(r)은 12V의 階段電壓에 對應되었는데, 이 f(r)에 對應되는 電壓을 y_i Channel Pulse로서 分離하여 特徵 Pattern을 原位置에 分配하는 過程을 거친다. 그러기 爲해서는 다음 (1)式을 定義한다.

$$y_i(r)=f(r)y_i \qquad (1)$$

$i=1,2,\cdots, N$. r=ㄱ, ㄴ, ㄷ, …, ㅎ, 여기서 f(r)은 번역된 임이의 文字이고, y_i는 Channel 선택 Pulse, $y_i(r)$은 그림 2의 一次標本 Pattern의 level A_i, B_i, $C_i, \cdots, Y_i$등에 對應된다. 따라서 ㄱ의 경우는

$$\begin{aligned} A_1&=f(ㄱ)y_1 \\ A_2&=f(ㄱ)y_2 \\ &\vdots \\ A_5&=f(ㄱ)y_5 \end{aligned} \qquad (2)$$

또 ㄴ에 대한 경우는

$$\begin{aligned} B_1&=f(ㄴ)y_1 \\ B_2&=f(ㄴ)y_2 \\ &\vdots \\ B_5&=f(ㄴ)y_5 \end{aligned} \qquad (3)$$

이와 같이 다른 모든文字에 대해서도 같은 方法이 적용되며, 선택된 文字의 Y Channel 分離가 이루어 진다.

Channel pulse y_1, y_2……y_5을 實現하기 위해서는 3bit Counter로 부터 分周矩形波를 發生시켜서 이것을 Matrix*에 加하여 出力端으로부터 각각 12V의 單一 Pulse y_1, y_2…… y_5를 發生시킨다. 이때 發生된 Pulse에는 分周波를 組合할 때 時差로 因하여 發生하는 不要 Pulse는 T_r 增幅 Clipper로서 除去하였다. 이 Channel pulse는 循還的으로 出力端에 나타나며 發生된 Pulse y_1, y_2, ……y_5는 그림 7에 표시하였다.

前記 번역된 文字 f(r)와 이 Pulse와의 And gate에 의하여 文字 Pattern이 선택된다.

이들 Pulse는 (1) (2)式의 $y_i(r)$ 또는 $A_1, A_2, \cdots A_5$ 및 B_1, B_2, ……, B_5, ……등에 對應되며, 이는 또한 그림6의 特徵의 部分 Pattern $\omega_1, \omega_2, \cdots \omega_5$ 및 그림 3의 標本 Pattern A_i, B_i, C_i, …… …, R_i등에 對應된다. 이들의 관계를 그림 6의

* 設計圖는 A—5에 주었다

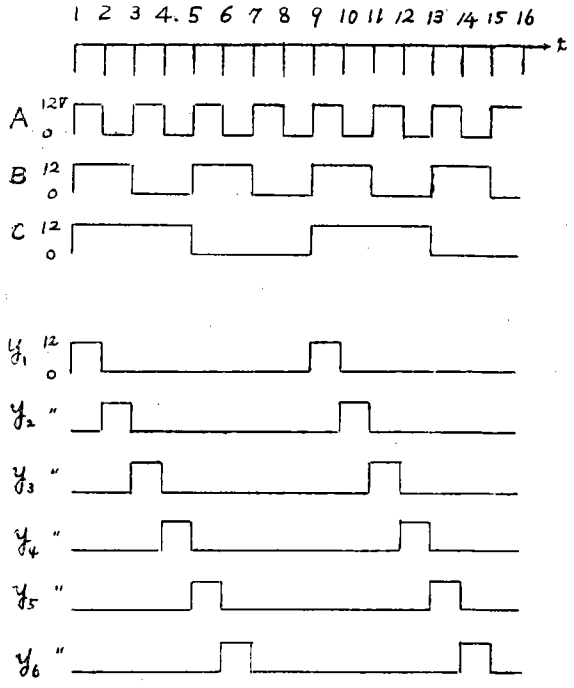


그림 7. YChannel Pulses.

特徵 Pattern과 각 文字의 標本 Pattern과를 비교하여, 同一位置에 있는 行의 部分 Pattern 集合 ω_i의 集計를 하면 A_i, B_i, C_i, ……, R_i를 論理變數로 하는 部分 Pattern 集合 ω_i가 (4)式과 같이 求하여 진다.

$(A_i, B_i, C_i, \cdots\cdots, R_i) \subset \omega_i$,

文字 ㄱ에 대해서는

$$\omega_1 = A_1$$
$$\omega_2 = A_5$$
$$\omega = \overset{3}{\underset{i=1}{U}} A_i \qquad (4)$$
$$\omega_4 = A_4$$

이 方程式은 文字의 標本 Pattern 그림 2의 A_i와 特徵 Pattern을 表現한 그림 6과의 關係를 나타내고 있는데 이 式으로부터 A_i를 Parameter로 하는 論理 Matrix가 쉽게 構成되어 진다. 다음은 行의 部分 Pattern ω_i를 論理變數로 하고 그림 6의 特徵 Pattern을 관측하면서 3個의 區分으로 類別된 部分 Pattern 集合 Ω_i의 列의 集計를 하면 文字의 特徵 Parameter가 順次的으로 出現하는 順序論理函數가 (5)式으로 주어진다. (文字 ㄱ에 대한 경우), $\omega_i \subset \Omega_i$ 즉

$$\Omega_1(\text{ㄱ}) = \overset{3}{\underset{i=1}{U}} \omega_i$$
$$\Omega_2(\text{ㄱ}) = \underset{i=1,4}{U} \omega_i \qquad (5)$$
$$\Omega_3(\text{ㄱ}) = \underset{i=1,3}{U} \omega_i$$

ω_i를 Parameter로 한 論理 System을 구성하면 文字의 特徵이 抽出된다.

다른 文字에 대해서도 같은 方法이 적용되는데 ㄴ의 경우는

$$\omega_1 = B_5 \qquad \Omega_1(\text{ㄴ}) = \overset{3}{\underset{i=1}{U}} \omega_i \qquad (6)$$
$$\omega_2 = \overset{4}{\underset{i=2}{U}} B_i \qquad \Omega_2(\text{ㄴ}) = \Omega_3(\text{ㄴ}) = \omega_1$$

ㄷ에서는

$$\omega_1 = \underset{i=1,5}{U} C_i \qquad \Omega_1(\text{ㄷ}) = \overset{3}{\underset{i=1}{U}} \omega_i \qquad (7)$$
$$\omega_2 = \overset{4}{\underset{i=2}{U}} C_i, \qquad \Omega_2(\text{ㄷ}) = \Omega_3(\text{ㄷ}) = \omega_1$$

ㄹ에서는

$$\omega_1 = \underset{i=1,3,5}{U} D_i \qquad \Omega_1(\text{ㄹ}) = \underset{i=1,4}{U} \omega_i$$
$$\omega_2 = D_4 \qquad \Omega_2(\text{ㄹ}) = \omega_1 \qquad (8)$$
$$\omega_3 = D_2, \qquad \Omega_3(\text{ㄹ}) = \overset{3}{\underset{i=1}{U}} \omega_i$$

ㅁ에서는

$$\omega_1 = \underset{i=1,5}{U} E_i \qquad \Omega_1(\text{ㅁ}) = \Omega_3(\text{ㅁ}) = \underset{i=1,5}{U} \omega_i$$
$$\omega_5 = \overset{4}{\underset{i=2}{U}} E_i \qquad \Omega_2(\text{ㅁ}) = \omega_1 \qquad (9)$$

ㅂ에서는

$$\omega_1 = \underset{i=3,5}{U} F_i \qquad \Omega_1(\text{ㅂ}) = \Omega_3(\text{ㅂ}) = \underset{i=1,5}{U} \omega_i$$
$$\omega_5 = \underset{i=1,2,4}{U} F_i \qquad \Omega_2(\text{ㅂ}) = \omega_1 \qquad (10)$$

이외의 모든 文字에 대해서도 같은 方法이 적용된다. 그런데 (4)~(10)式을 관찰하면 ㄱ의 경우는 部分 論理函數가 相異하지만, ㄴ, ㄷ, ㅁ, ㅂ의 경우는 $\Omega_2 = \Omega_3$, $\Omega_1 = \Omega_3$와 ω_i에서 B_i C_j, ……, R_i등이 同一函數가 나타난다. 다른 文字에서도 이와 같은 函數의 關係가 多數 抽出된다는 것이 極히 注目된다. 이는 한글 文字의 獨特한 구조라고 볼 수 있다. 따라서 이를 實現하는데 가장 경제적이고, 간결한 回路의 設計는 論理方式이란 것이 明白해 진다. 文字當 15bit의

情報量을 抽出해 낼 때에는 素子數가 3分之 1의 節約을 가져왔고 (다른 識別函數에 비해서). 5×7meoh=35bit의 경우는 約 5分之 3의 節約을 가져왔다.

이는 大端히 興味있는 問題라 보며, 英文字등에서는 文字의 情報量을 增加시키면 素子數는 比例하는데, 한글文字에서는 論理函數를 적용할 때에는 逆比例한다. 本 硏究에서는 特異한 利點을 着目하였으며, 한글文字에 관한 限 特徵抽出部에 論理函數의 적용은 莫大한 素子의 節約을 가져 온다는 確證을 얻었다.

이와 같은 論證에 입각하여 回路論計는 (4)~(10)式으로 부터 A_i, B_i…, R_i를 論理變數로 하는 24個의 群으로 구성되여 지는 特徵分配기구와 여기서 抽出된 部分 Pattern ω_1, ω_2, …, ω_5를 2次 論理變數로하는 論理기구*를 구성하였다.

그의 出力端으로 부터 原 Patter配列과 同一한 文字 Pattern의 集合이 형성되어서 所定의 特徵 Parameter가 抽出된다.

이때 文字의 出現過程이 그림 8 (b), (c), (d)와 같은 順序로 나타나면 未知의 入力文字는 ㄱ라는 判定이 내려진다. 이 出力端에 表示器를 연결하면 原文字가 表示된다. Coding하여 電子計算機에 直結할 수 있고, 또 천孔기에 걸면 文字 Code가 천孔되며, 記憶장치에 걸면 文字 Pattern이 記憶되어 永久保存이 可能하다. 表示器도 위의 5個文字에 대해서는 完成하였으나, 本紙는 特徵抽出까지 끝맺고, 제한된 紙面上 다음 機會에 發表예정이다.

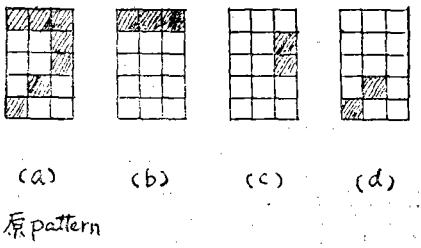


그림 8. Letter Elements의 出現順序
Fig. 8 The Appearance Sequence of Letter Elements.

* 設計圖 A—6(Ⅰ), (Ⅱ), (Ⅲ)에 주어졌다.

Ⅴ. 實驗結果

標本에 대한 特徵抽出의 實驗과 그 外의 몇가지 수반되는 問題點에 대해서 Test 하였다.

(1) 이 System이 垂直으로 配列된 Cell로부터 檢出된 理想的特徵을 認識할 수 있는 能力이 있는가의 與否를 決定짓기 爲하여 Test하였다. 5個文字의 標本 Pattern을 順次的으로 記憶 Register에 加하여 認識 Matrix의 감시 Lamp로서 관측한 즉 10回의 試驗에서 한번도 誤判定이 없었다. 또 特徵抽出回路의 性能을 조사하기 爲하여 認識 System과 結合하여 同一試驗을 반복한 結果 100% 기능을 發揮하였다. 따라서 電子기구部分은 더 改良의 必要없이 完全하다는 것을 確認하였다. 그러나 初段驅動기구의 기계적 部分은 아직 不安全한 경우가 나타나서 좀더 改善이 必要하다.

3) CRO에 의한 測定値와 비교하기 위하여 特徵出端에 出現하는 特徵抽出 Ω_1, Ω_2, Ω_3의 出現位置를 X軸을 時間으로 하고, Y軸을 Ω로하면 理論的인 Pattern (ㄱ의)分布는 그림 9와 같이 된다.

이는 實際와는 90度回轉시켜서 그린 것인데 CRO의 測定値와 一致시키기 위해서 회전시켜서 그린 것이다. 그 理由는 文字 Pattrn의 Channel 分離時, Ychannel Pulse로서 分離했기 때문에 이는 CRO의 時間軸과는 꼭 90度의 位相差가 생기기 때문이다. 表示器를 사용하면 그냥 定常値가 된다.

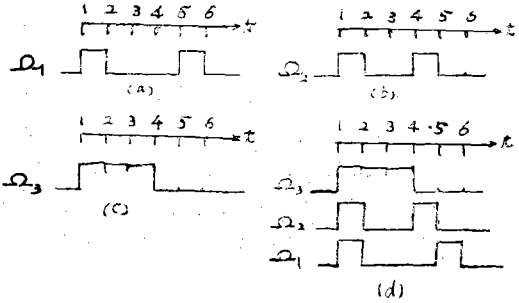
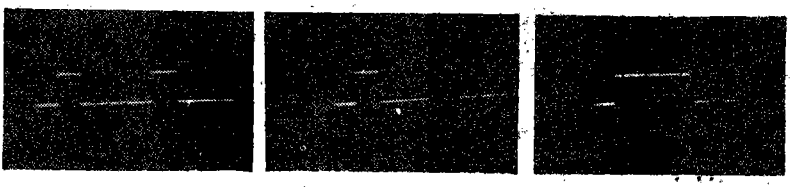
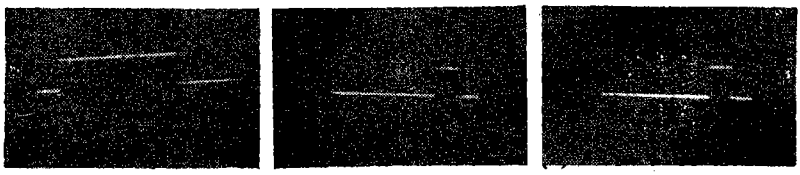


그림 9



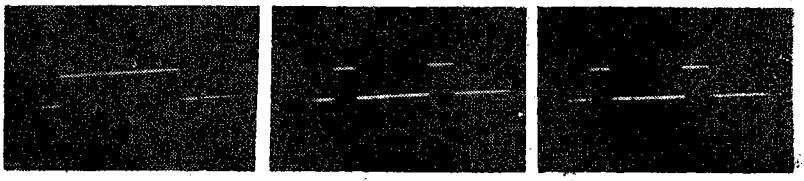
Ω1 Ω1 Ω3

ㄱ



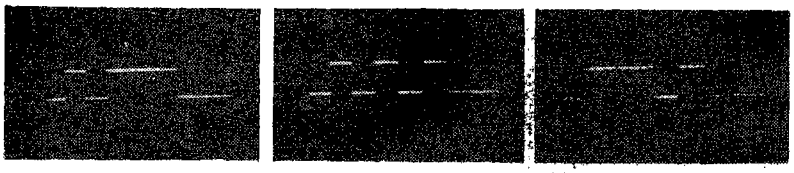
Ω1 Ω2 Ω3

ㄴ



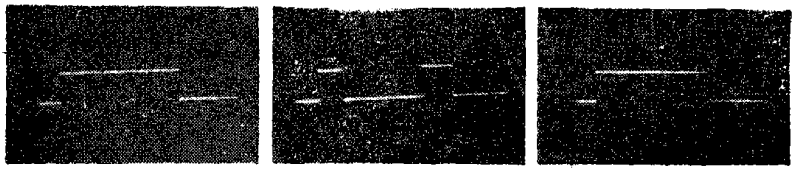
Ω1 Ω3 Ω3

ㄷ



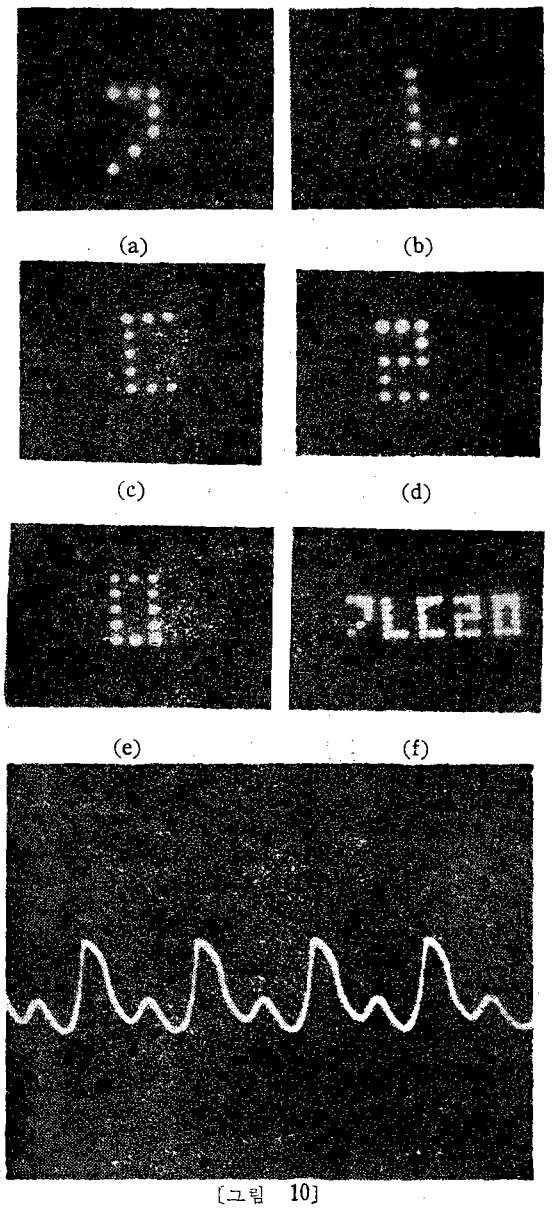
Ω3 Ω2 Ω1

ㄹ



Ω1 Ω3 Ω2

ㅁ



〔그림 10〕

그림 9(d)는 (a), (b), (c)를 個別的으로 檢出해서 時間軸에 마추어서 그린 것이다.

4) 그림 10은 CRO로서 Ω_1, Ω_2, Ω_3을 獨立的으로 관측한 것이다. 그림 9에서의 경우와 같이 Ω_1, Ω_2, Ω_3를 時刻點에 마추면 각 文字의 Pattern 位置가 보인다. 勿論 이때 이들을 90度 時計方向으로 돌려서 보아야 한다.

그림 10은 ㄱ, ㄴ, ㄷ, ㄹ, ㅁ의 5個母音에 대한 抽出 Pattern이다.

CRO로서 관측할 수 있는 것은 Ychannel pulse가 Cyclic로 동작하게 設計되었기 때문에 CRO관측이 可能하다.

5) 文字表示의 한 方法으로서 一般 CRO인 Du Mont 304-A로서 揮度變調를 하여 文字表示를 시도하여 보았다.

그림 11과 같이 滿足한 結果를 얻었다. 여기서 文字 Pattern數가 적은것 같은데 이는 장치의素子數를 最小로 한 豫定된 것으로, 文字認別에는 아무 지장이 없다.

6. 每秒認識할 수 있는 速度를 測定하기 위하여 文字紙를 高速度로 移動시켰더니 이 System이每秒數十字의 비율로서 認識할 수 있음을 알게됐다. 그러나 光檢出 驅動部의 기계적인 速度제한과 電動 Typwriter를 쓴다고 하는 가정에서는 每秒 5字程度로 쓰일 것으로 본다.

7) 文字의 경사의 効果를 결정하기 위하여 文字를 左右로 約 5°가량씩 경사시켜서 認識한 結果 이때의 認識率은 約 70%였다. 文字가 走査할 때 光檢出部에 反射鏡을 달아서 文字를 擴大시켜서 檢出하여야 할 것이므로 文字는 이상적인 상태에서는 指定된 位置에 오지만 一般으로 回轉, 경사를 고려하는 것이 認識기구가 그만큼 여유가 있을 것이며 이 實驗은 우리 日常生活에 筆記體를 主로 쓰는 사실을 감안한다면 더욱 重要하다고 보겠다. 現在 約 5°정도 경사되어도 별 지장이 없다,

이外에도 다음 硏究의 資料를 얻기 위하여 여러가지 實驗을 할 계획으로 있다.

Ⅵ. 結 論

1). 한글文字의 모아쓰기 認識도 같은 方法으로 實現되지만 24子母를 識別할 때를 비한다면 約 400倍의 素子數가 必要하게 된다.

따라서 이 分野가 發達되면 더 좋은 認識方法이 나오지 않은 限 풀어쓰기가 實用化될 可能性이 많다.

2). 이 연구의 結果로부터 한글文字의 變化없는 單調로운 缺點이 縱橫으로 되어 있기 때문에 同一한 識別函數가 多數 유도되어 論理구성에서

는 큰 長點으로 나타난다는 것을 알았다.

3). 特徵抽出 System의 設計에서 多方面으로 最少 素子의 設計를 위한 검토를 하여 많은 材料의 節約을 보았다.

4). 이 硏究過程에서 얻어진 모든 資料들은 이 分野의 硏究에 약간의 기초가 된다고 보겠다. 音聲 Typwriter, 文字의 傳送, 末端學習기계등 이에 관련된 硏究에 있어서 한글文字의 認識 Pattern에 대한 실마리가 잡혔다고 보겠다. 끝으로 이 硏究는 오랜 時日을 경제적 뒷받침이 없어서 實驗을 못하다가 今番東亞日報社에서 후원하여 結實을 보게 된것을 感謝하며, 격려하여 주신 電氣科敎授님들과 이 硏究에 助力하여준 四學年 李均夏, 全澤, 李容圭등에게 深深한感謝를 드리는 바이다.

參 考 文 獻

1. F. Rosenblatt; The perceptron, psychological Review, 65.8.1958. P386~408
2. W.H. Highleyman; Linear Decision Functions, with Application to pattern Recognition. IRE. Vol.50 No.6 P1501~1514. June, 1962
3. L.F. Turner; A System for the Automatic Recognition of Moving patterns. IEEE, Vol. IT-12, No.2, P195~205, April. 1966.
4. UDAKAWA. et; A parallel Two-stage Decision method for Statistical character Recognition and A Method for Estimation of positional Distribution of character. J.IECE. Japan. 48.9 Sept. 1965.
5. N.J. Nilsson; Learning Machines. Foundations of trainable pattern-classifying Systems. 1965. McGRAW-Hill co.
6. J.T. Tou; Application Automata theory. Academic. Press 1968.
7. J.T. Tou; Computer and information Sciences-11 Academic Press. 1967.
8. J.T. Tou; Digital and Sampled-Data Control Systems. McGRAW-Hill Co. 1959.
9. Yaohan chu; Digital computer Design Fundamentals. McGRAW-Fill Co. 1962.
10. 坂井利之：情報處理とその裝置. 1967. 共立社