OCR 貼上池「其他格式」按空白分段

STK32 (Harbour) + img2text 2026-06-24

一、訴求與現象

在 GET 欄位 Ctrl+V 選「其他格式」做一般 OCR，貼上池(雙擊選取的清單)會把同一行的多欄文字黏成一長串。例如尺寸對比圖兩個料號 1614255680 與 329333 變成「1614255680329333」、「小孔 25 mm」與「小孔 26 mm」黏在一起。需求：同一行應按空白分成多段文字，分別放入貼上池。

二、根因

img2text.py 讀 System.ini [OCR] 的 StripSpace（預設 1），會呼叫 strip_spaces 把每行所有空白（半形 / Tab / 全形）全部移除，本意是讓單值貼上乾淨（如把 1614 255 680 併成 1614255680），但用在貼上池時反而讓相鄰兩欄黏連。貼上池端原本只按換行切，每行一段，所以呈現一長串。

三、需求演進與設計選擇

階段	決策
v1	同行按空白切成多段（每個 token 一項）
v2	修正：單一空白不分，連續兩個以上空白才分段（避免「小孔 25 mm」被切碎）
v3	規則併入 System.ini [OCR]，做成可調門檻；分段邏輯放 img2text（執行期讀 ini，改門檻免重編）

四、踩坑與陷阱

陷阱	說明
StripSpace 雙面刃	去空白對「單值貼上」有益，但對「貼上池多欄」有害；不能直接全域關閉，需隔離。
分段該放哪一端	放 PRG 需 Stk32 重編才能調門檻；放 img2text 則執行期讀 [OCR]，改 ini 即生效，且 PRG 維持單純。

陷阱	說明
bash heredoc 吃跳脫	用 bash heredoc 產生含反斜線 n 的 Python 程式碼時，字串內的反斜線 n 被吃成真換行，導致 img2text.py 出現 unterminated string literal。改用 Write 工具寫修正腳本、以 bytes([92,110]) 明確寫字面，並用 py_compile 驗證。

五、解決方法

檔案	修正
IMG2TXT/img2text.py	新增 CLI --keepspace：該次跳過 strip_spaces，並依 [OCR] PoolSplitSpaces(預設2) 把「連續 N 個以上空白(先全形轉半形)」用 re.sub 轉成換行做欄位分段；單一空白保留同段。重編 img2text.exe。
HBSTK/source/TLIB.PRG	OcrClipToText(cDoc, lKeepSpace) 加參數，lKeepSpace 時命令列帶 --keepspace；PasteHistoryPick「其他格式」呼叫帶 .T.；貼上池維持按換行分段（分段交由 img2text 決定）。
System.ini [OCR]	新增 PoolSplitSpaces=2（執行期可調，免重編）。
ini_editor.py + ini_template.json×3	同步新增 PoolSplitSpaces（標籤「貼上池分段空白數」，可選 2/3/0）。

其他 OCR 用途（行照 / 名片用 --doc、單值貼上 PasteImageAsText）維持 StripSpace=1，完全不受影響。

六、使用技術與用法

技術	用法
CLI 旗標隔離行為	以 --keepspace 單次覆蓋 ini 設定，只對貼上池路徑生效，不動全域 StripSpace。
正規表示式分段	re.sub(r" {N,}", 換行) 把連續 N 個以上空白轉成換行，達成「欄位分段、單空白保留」。
設定外部化	門檻寫在 System.ini [OCR] PoolSplitSpaces，img2text 執行期讀取，調整免重編 Stk32。

技術	用法
設定三同步	System.ini 值 + ini_editor.py INI_META + ini_template.json×3，新增 key 一律三處同步。
產碼工具選擇	需要產生「含反斜線跳脫」的程式碼時，避免 bash heredoc，改用 Write 工具或明確 byte 值，再 py_compile 驗證。

七、效果

門檻=2 時：兩料號之間的大間隔（連續多空白）會分段成 1614255680 與 329333 兩項；「小孔 25
mm」因只有單一空白而保留為一項。日後要更細或更粗，只需改 System.ini 的 PoolSplitSpaces。