

変動のある様式を対象とした文字情報抽出方式

東屋那音[†] 工藤司[†]

静岡理科大学 情報学部[†]

1. はじめに

光学文字認識は、定型様式や文字のみの文書では高い認識率を達成している。しかし、ガントチャートのように、文字と図形の要素が混在し位置も変動する場合には、文字情報の抽出精度が低下するという課題がある[1]。さらに、付帯する日付や作業情報と関連付けるため、図形の位置情報も抽出する必要がある。

そこで本研究では、まず図形の位置を認識し、その中の文字情報を抽出すると共に、認識した位置情報に基づき、図形と付帯する情報を関連付ける方式を提案する。さらに、実際の工場のスケジュール表を対象として、ガントチャートのような変動のある様式からも、効率的に工程情報を抽出することが可能であることを示す。

2. 本研究が対象とする様式

本研究が対象とするスケジュール表の例を図1に示す。なお、対象工場では現在、このスケジュール表から工程情報を手入力しており、本研究はこの入力の自動化を狙ったものである。図1に示す様に、上部に月日情報、左に対象名の情報、中央に作業内容と工程が四角で示される。例えば、図1の太線の四角で示された工程の場合、開始月日が8月3日で、対象名がA1となる。また、文字情報を読み取る際には、ノイズとなる破線の除去を行う必要がある。

8月	1	2	3	4	5	6
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	
日	月	火	水	木	金	土
A1			RXT555NE6-01(野・原) SHIN666 静岡部品(長野) 加工			
B1		RXT555NE6-02 SHIN777 SHIZUOKA(山梨) 加工		RXT555NE6-06 NOSUKE777 HUKUROI(富士) 金メッキ		

図1: スケジュール表

3. 提案方式

2節の様式に対し、以下の手順で情報を抽出する方式を提案する。

(1) 枠線の抽出

スケジュール表から作業対象を抽出するため、図2①の太線に示すように、一定以上の長さの線分を抽出する。

Character Information Extraction Method from Non-standard Form

[†]Naoto HIGASHIYA [†]Tsukasa KUDO

[†]Shizuoka Institute of Science and Technology

(2) 日付と対象名の抽出

(1)の最上段で抽出した日付欄に対し、あらかじめ準備した各日付の画像とのパターンマッチングを行い、図2②に示す日付と、日付を含む枠の位置を抽出する。また、①で抽出した線から輪郭抽出を行い、対象名とその枠の位置を抽出する。対象名の順番は固定であるため、位置の情報に基づき、個々の対象名を決定する。日付の月(8月)もこの時点で抽出する。

(3) 文字の読み取り

破線は、直線として位置を抽出し、画像修復を利用して図2③のように取り除く。その際、工程の枠の位置を特定する。その後、工程ごとに文字の読み取りを行い、(2)で抽出した日付や対象との関連付けを行う。

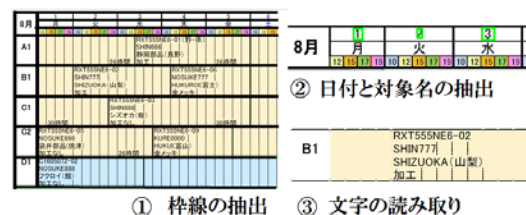


図2: 画像の抽出手順

4. 実装

提案方式を、対象工場の生産管理システムへの入力を前提とした入力機能として実装した。具体的には、スキャンした図1の画像ファイル(PNG形式)から自動的に表1の工程情報を抽出し、入力用のテキストファイルを作成した。

表1: 抽出した工程情報の事例

月	日付	工程情報	対象名
8	3	RXT555NE6-01(野・原) SHIN666 静岡部品(長野) 加工 26時間	A1

開発言語にはpython3.7.1を利用した。また、図形の抽出は画像処理ライブラリOpenCV4.0.0を使用し、表2の関数を使用して実装した。また、OCRライブラリは交換が可能な構成とし、比較評価のためTesseract-ocrとGoogle Vision APIを使用した。文字認識は、図1の太枠の工程の左側に示される「26時間」の様な断片の文字を除くため、認識した文字数が10文字以上の枠のみを工程の対象とした。

表2: OpenCVの使用関数

実装機能	関数名
枠線抽出, 破線認識	HoughLinesP
四角形抽出	findContours
日付抽出	matchTemplate
画像修復	inpaint

5. 実験と評価

文字の読み取り精度と入力効率を評価した。事前に実施した予備実験で、図1のスケジュール表全体(1ヶ月分)を一括で処理すると、OCRの精度が劣化した。

そのため、本実験では図1のように1ヶ月のスケジュール表を6日ごとに分割し、最初の6日を4ヶ月分、合計16工程を対象とした。

(1) OCRライブラリの精度比較

4節のOCRライブラリの精度を比較評価した。この結果、図3に示すようにカタカナで大きな差異が見られた。また、文字ごとにも、表3に示すようにライブラリの間で認識可否が異なった。

表3: 文字ごとの精度

正しい文字	ル	ヤ	イ	自	寒
Google Vision API	11	ヤ	1	自	V m H I
Tesseract	ル	ヤ	イ	目	ミコ

(2) 工程ごとの抽出精度比較

工程ごとに文字の認識および破線削除の精度を評価した。OCRライブラリはTesseract-ocrを使用した。この結果、図4に示すように、工程ごとに精度のばらつきが見られた。文字認識のエラーは、図3に示すように漢字、カタカナが多く、平均では約4%であった。破線削除のエラーは、文字と同じ大きさの破線が削除できない割合が高く、平均では、同じく約4%となった。なお、日付と対象の関連付は正しく行われた。

(3) 入力・確認時間の比較

工程、日付、対象名を手入力した場合と、本方式で入力した場合の効率を比較評価した。どちらの場合も入力後にデータを確認し、エラーがあれば修正した。

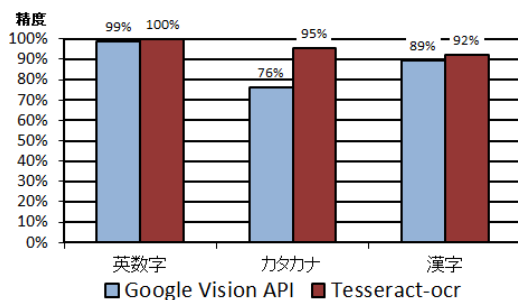


図3: OCRの精度比較

エラーの数は、手入力の2件に対し、本方式では56件であった。従って、図5の「確認時間」に示すように、本方式では修正に時間を要した。しかし、入力自動化されるため、1件当たりの入力・確認時間の平均は、従来の約98秒に対し27秒になり71秒短縮された。この結果、全ての月で入力・確認時間が65%以上短縮できた。

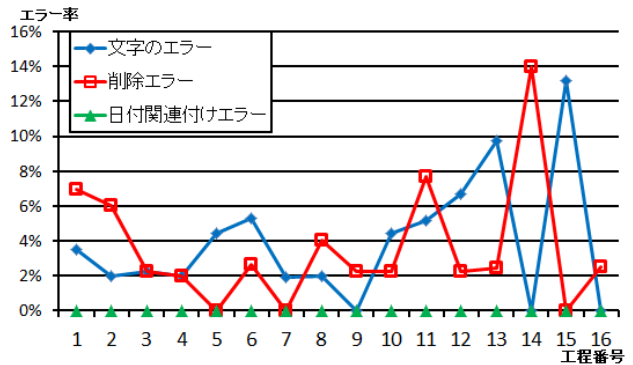


図4: 工程別エラー率

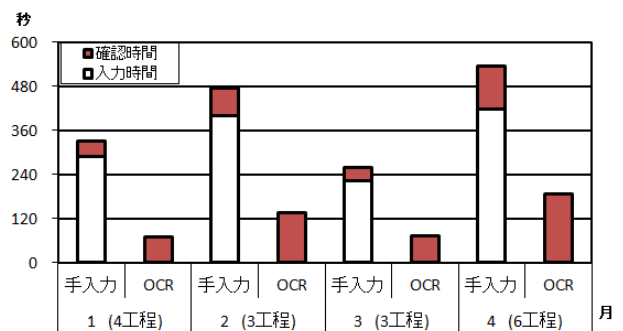


図5: 入力・確認時間

6. 考察とまとめ

本稿では、図形が混在する画像から文字情報を抽出する方式を提案し、ガントチャートからも工程の文字情報と付帯情報が抽出できることを示した。さらに、実際のスケジュール表を対象とした実験により、手入力と比較してエラー率は高かったものの、入力を自動化できるため全体としては効率を向上できることを示した。

また、入力精度に関しては、OCRライブラリによりエラー発生個所にばらつきがあることが分かった。従って、2つのライブラリを併用して一定のエラーを自動検出することで、確認時間をさらに短縮できると考えられる。

本研究はJSPS科研費19K11985、および東海産業技術振興財団の助成を受けたものです。

参考文献

[1] 上條治夫: 視覚障害者による化学構造式認識法の検討, 化学と教育, 64(6), pp. 272-275 (2016).