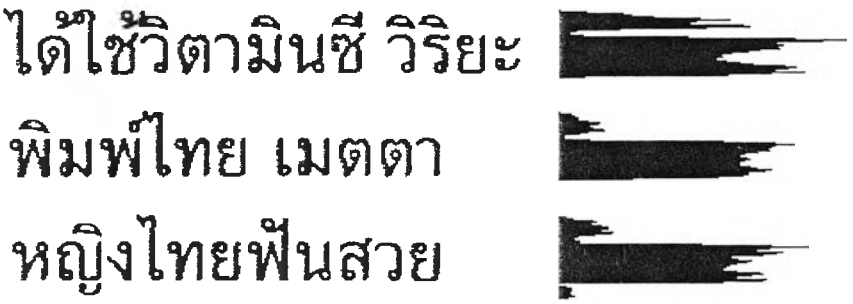
บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

แนวคิดและทฤษฎี

การตัดแยกตัวอักษรที่ติดกัน(Segmentation of touching character) เป็นส่วนหนึ่งในการรู้จำตัวอักษร (Character Recognition) ซึ่งเป็นการนำภาพตัวอักษรมาทำการตรวจสอบ ว่าเป็นอักษรที่ติดกันหรือไม่ แล้วทำการตัดแยกให้เป็นตัวอักษรเดี่ยว เพื่อส่งต่อไปให้ส่วนรู้จำตัวอักษรเดี่ยวทำการวิเคราะห์ให้ได้เป็นตัวอักษรแบบเท็กซ์ต่อไป การตัดแยกตัวอักษรสามารถนำทฤษฎีต่างๆ มาประยุกต์ใช้งานดังต่อไปนี้

1. การหาแนวตัดแยกบรรทัด และคอลัมน์โดยอาศัยโปรเจกชั่น [5] [7]



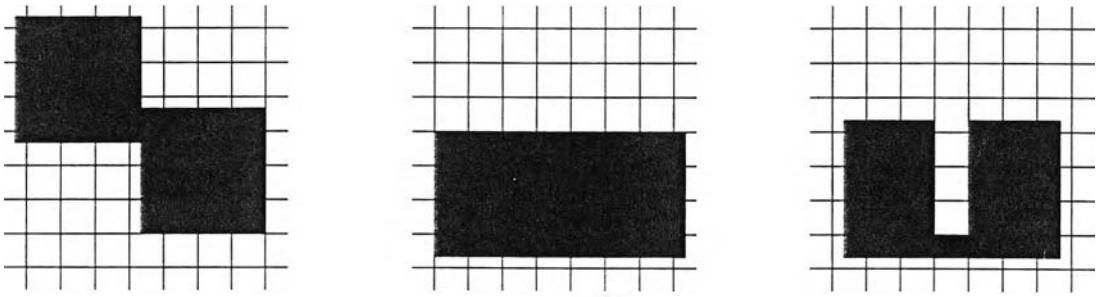
รูปที่ 2-1 แสดงโปรเจกชั่นแนวนอน

การหาโปรเจกชั่นของภาพในแนวนอน(Horizontal Projection) เมื่อค่าโปรเจกชั่นจุดใดที่มีการเปลี่ยนจาก 0 เป็นมีค่า ใดๆ ก็ตามให้สันนิษฐานว่าเป็นจุดเริ่มต้นของเส้นบรรทัดบน ถ้าเส้นระหว่างบรรทัดห่างพอ กล่าวคือ ค่าของโปรเจกชั่นในแนวนอนมีค่าจะน้อยลงจนเท่ากับ 0 อีกครั้ง จะทำให้สามารถตัดแบ่งออกเป็นสองบรรทัดได้ โดยทำตัดแยกตรงค่าดังกล่าว จากนั้นทำการแยกส่วนของบรรทัดออกมาทีละบรรทัด จนหมด

รูปที่ 2-2 แสดงโปรเจกชั่นแนวตั้ง

โปรเจกชั่นในแนวตั้ง(Vertical Projection) พิจารณาเช่นเดียวกันกับโปรเจกชั่นในแนวนอน ซึ่งจะทำให้สามารถตัดตัวอักษรจากบรรทัดออกเป็นตัวอักษรทีละตัว(Character Segmentation) แต่ในกรณีที่ตัวอักษรติดกันก็จะไม่สามารถแยกออกได้เนื่องจากค่าของโปรเจกชั่นกรณีของตัวอักษรที่ติดกันไม่ได้เท่ากับ 0 ฉะนั้นจะได้กลุ่มของภาพอักษรที่ติดกัน

2. การหาแนวตัดแยกตัวอักษรที่ติดกันโดยใช้นิวเมตริกซ์ [8]



(a) ภาพ (b) โปรเจกชั่น (c) นิวเมตริกซ์

รูปที่ 2-3 นิวเมตริกซ์สำหรับตัดตัวอักษรที่ติดกัน

การหาแนวตัด(Break Cost) [8] โดยอาศัยนิวเมตริกซ์ จากรูปที่ 2-3 กำหนดให้ภาพ (a) เป็นส่วนของภาพตัวอักษรที่ติดกัน ภาพ (b) เป็นค่าที่หาได้จากฟังก์โปรเจกชั่นในแนวดิ่ง และ ภาพ (c) เป็นค่าที่ได้จากการคำนวณแบบนิวเมตริกซ์ในแนวดิ่ง ซึ่งสามารถคำนวณได้คล้ายกับการหาโปรเจกชั่นในแนวดิ่งของจุดภาพ แต่เพิ่มการคำนวณทางตรรกศาสตร์เข้ามา โดยการ

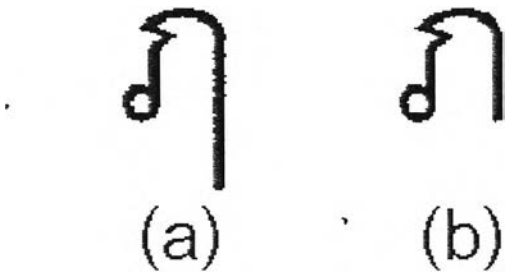
ใช้ตรรกศาสตร์ "และ"(AND) กับจุดดำข้างๆ ในแนวนอนก่อนทำการหาค่าโปรเจกชั่นในแนวดิ่ง บริเวณที่มีค่านิวเมตริกซ์น้อยที่สุดจะกำหนดให้เป็นจุดตัด ซึ่งการหาแนวตัดแยกโดยวิธีนิวเมตริกซ์นั้นสามารถทำได้ทั้งในแนวนอน และแนวดิ่งเช่นเดียวกับการหาแนวตัดแยกโดยวิธีโปรเจกชั่น

3. การหาแนวตัดแยกตัวอักษรที่ติดกันโดยหาจากอัตราส่วนระหว่างอนุพันธ์อันดับ 2 ของ โปรเจกชั่น กับค่าของ โปรเจกชั่น [9]

หาแนวตัดแยกของตัวอักษรที่ติดกัน 2 ตัว จุดเชื่อมของ 2 ตัวอักษรจะมีค่าของ Vertical Projection(V(x)) เปลี่ยนแบบ Sharp Minimum ซึ่งแสดงถึงบริเวณที่เกิดตัวอักษรติดกัน เนื่องจากมีปริมาณจุดภาพน้อย ดังนั้นบริเวณที่เป็นจุดตัดหาได้จากอัตราส่วนระหว่าง อนุพันธ์อันดับ 2 คือ V(x-1) – 2V(x) + V(x+1) กับค่าของ Projection

$$Cut\_position = \frac{V(x-1) - 2V(x) + V(x+1)}{V(x)}$$

4. ทฤษฎีที่เกี่ยวกับลักษณะบ่งความต่าง เป็นเครื่องมือในการแยกแยะสมาชิกของกลุ่มใดๆ ออกจากกัน โดยสามารถใช้ได้กับการแยกแยะสมาชิกของกลุ่มใดๆก็ได้เพียงอาศัยความแตกต่างและความคล้ายกันของสมาชิกในการแบ่งกลุ่ม [16] Ogden นำเสนอทฤษฎีปฏิลักษณ์ ซึ่งเป็นพื้นฐานที่ช่วยให้การบ่งความต่างมีประสิทธิภาพ โดยได้แบ่งประเภทของปฏิลักษณ์ออกเป็น 3 อย่างคือ แบ่งโดยการตัด (cut) แบ่งโดยปริมาณ (scale) แบ่งโดยนิยาม (Definition) เช่น มี-ไม่มี ดังรูปที่ 2-4 แสดงลักษณะบ่งความต่างของตัวอักษรโดยภาพ (a) มีส่วนหาง ภาพ (b) ไม่มีส่วนหาง ดังนั้นจึงสามารถทำการแบ่งกลุ่มของตัวอักษรที่อยู่ต่างระดับได้โดยอาศัยการตรวจสอบกลุ่มของจุดดำบริเวณหางของภาพ 'ภ' และ 'ภ' แต่ในบางกรณีการแยกแยะสมาชิกของกลุ่มใดๆ ไม่สามารถทำได้โดยอาศัยลักษณะบ่งความแตกต่างเพียงอย่างเดียว จำเป็นต้องใช้คุณลักษณะมากกว่าหนึ่งลักษณะในการบ่งความต่างเช่น 'โ' กับ 'ป' ทั้งคู่มีส่วนของจุดภาพอยู่ในระดับที่ 1 และ 2 แต่ 'ป' จะมีความกว้างของจุดภาพมากกว่า 'โ'



รูปที่ 2-4 แสดงลักษณะบ่งความต่างของตัวอักษร

5. ตัวอักษรไทย

ตัวอักษรไทยมีวิวัฒนาการมาจากตัวอักษรอินเดียตอนใต้ ซึ่งแตกแขนงออกเป็นอักษรขอม และมอญ แต่เดิมพ่อขุนรามคำแหงมหาราชได้ดัดแปลงอักษรขอม และมอญให้เป็นอักษรไทยที่มีสระ พยัญชนะ และวรรณยุกต์ อยู่ในบรรทัดเดียวกัน ต่อมามีการเปลี่ยนแปลงให้สระอยู่ด้านหน้า ด้านหลัง ด้านบน หรือด้านล่าง และวรรณยุกต์อยู่ด้านบน ช่วง พ.ศ. 2538-2540 มีการปรับปรุงตัวพิมพ์ใหม่เป็นตัวว่ดหัวกลมเส้นบางเสมอกัน เส้นตั้งฉาก และแนวนอนของตัวอักษรเป็นระเบียบขึ้น ซึ่งเป็นลักษณะของ "ตัวเหลี่ยม" ในปัจจุบัน ในตอนแรกๆ ตัวพิมพ์จะมีลักษณะนี้ทั้งสิ้นไม่ว่าจะเป็นขนาดใดก็ตาม ในปี พ.ศ. 2477 มีหนังสือบางเล่มพิมพ์ด้วยตัวหนาสันนิษฐานว่าตัวพิมพ์แบบหนา หรือตัวโป้งจะเริ่มมีในระยะนี้ ในปี พ.ศ. 2547 มีตัวอักษรที่เรียกว่าตัวฝรั่งเศสเกิดขึ้น เป็นการเลียนแบบอักษรโรมันคือ เส้นมีความหนาบางต่างกัน หลังจากนั้นไม่นานมีการหล่อตัวพิมพ์จิ๋วขึ้นใช้ด้วย ในปี พ.ศ. 2468 มีการหล่อตัวพิมพ์ภาษาไทยขึ้นใช้หลายแบบหลายขนาด คือมีทั้งตัวเหลี่ยม ตัวฝรั่งเศส ตัวเอน ตัวจิ๋ว และได้มีการดัดแปลงแก้ไขรูปแบบการพิมพ์เรื่อยมา [5]

ตัวอักษรไทยจะประกอบด้วย ตัวอักษรพยัญชนะ สระ วรรณยุกต์ ซึ่งสามารถแยกออกให้เห็นดังนี้

1. พยัญชนะ มีทั้งหมด 44 ตัว ปัจจุบันใช้งานเพียง 42 ตัว อีก 2 ตัวไม่ได้ใช้งาน คือ ฃ ฅ
2. สระ สามารถแยกออกเป็น สระระดับบน ระดับล่าง และระดับเดียวกับพยัญชนะ
3. วรรณยุกต์ มีทั้งหมด 4 ตัว
4. ตัวเลข 10 ตัว

เพื่อความสะดวกในการตัดตัวอักษรไทยที่ติดกัน จึงแบ่งประเภทของตัวอักษรออกเป็นกลุ่มต่างๆ 5 กลุ่มตามระดับของตัวอักษรที่ปรากฏในบรรทัด ดังแสดงตามตารางที่ 2-1

ตารางที่ 2-1 แสดงการแบ่งกลุ่มตัวอักษรไทย

กลุ่มที่	เกณฑ์การแบ่ง	สมาชิก
1	อักษรที่อยู่ระดับหนึ่ง	ิ ี ึ ื ็ ่ ้ ๊ ๋ ั ์ ๆ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ ๐
2	อักษรที่อยู่ระดับสอง และสาม	ป ฝ ฟ ฬ ไ ใ โ
3	อักษรที่อยู่ระดับสาม	ฯ ะ เ า ก ข ค ฆ ง จ ฉ ช ฌ ฒ ฌ ณ ด ต ถ ท ธ น บ ผ พ ภ ม ย ร ล ว ศ ษ ส ห อ ฮ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ ๐
4	อักษรที่อยู่ระดับสาม และสี่	ฦ ฤ ฯ ฎ ฏ ญ ฐ
5	อักษรที่อยู่ระดับสี่	ุ ู

เอกสารและงานวิจัยที่เกี่ยวข้อง

Shuichi Tsujimoto and Haruo Asada [8] การตัดแยกภาพตัวอักษรนั้น มักมีปัญหาในการตัดแยก อันเนื่องมากจากความคุมเครือของตัวอักษรที่ต้องการจะตัดเช่น 'r' กับ 'n' เมื่อตัวอักษรติดกันแล้ว จะทำให้คล้ายกับตัวอักษร 'm' หรือ ในกรณีของตัวอักษร 'i' ที่มีกลุ่มของจุดภาพ 2 กลุ่มโดยที่ไม่มีส่วนใดของกลุ่มของจุดภาพทั้ง 2 กลุ่มสัมผัสกันเลยทำให้เกิดความผิดพลาดในการตัดแยกตัวอักษรได้ ดังนั้น Shuichi Tsujimoto และ Haruo Asada ได้นำเสนอวิธีการแก้ไขข้อผิดพลาดดังกล่าวโดยใช้ Decision Tree ในการหาจุดตัดที่เหมาะสม แต่พบว่าในบางครั้งการติดกันของตัวอักษรเช่น 'rm' อาจวิเคราะห์ได้เป็น 'r' 'm' กับ 'n' 'n' ดังนั้นจึงใช้ความรู้ทางด้านภาษามาช่วยในการตัดสินว่าควรจะใช้ตัวอักษรใด จากการทดสอบกับเอกสารจำนวน 32 ฉบับ พบว่ามีความถูกต้อง 99.7%

Su Liang, M. Ahmadi, M. Shridhar [10] นำเสนอวิธีการตัดแยกตัวอักษรที่ติดกันโดยใช้ Discrimination Function และ Dynamic recursive segmentation algorithm เริ่มจากการตรวจหาบริเวณที่ติดกัน แล้วหาจุดตัดแยกโดยการหาค่าอนุพันธ์อันดับสองของค่าโปรเจกชั่นในแนวดิ่ง กับค่าของโปรเจกชั่นในแนวดิ่งเพื่อทำการตัดแยก จากนั้นตรวจสอบส่วนที่ตัดออกมาว่าใกล้เคียงกับ pattern ที่เตรียมไว้หรือไม่ ถ้าไม่ก็นำตัวที่ตัดมารวมกับกลับเป็นตัวเดิมก่อนตัด ทำ

การทดสอบกับเอกสารจำนวน 12 ฉบับ พบว่ามีความถูกต้องสูงสุด 99.85% โดยมีตัวอักษรที่ติดกัน 40% และความถูกต้องต่ำสุด 99.4% มีตัวอักษรที่ติดกันประมาณ 60%

งานวิจัยของ วิชา พานิช [1] ทำการรู้จำตัวอักษรพิมพ์ไทยโดยใช้ลักษณะบ่งความต่างของตัวอักษรภาษาไทย ได้แบ่งงานออกเป็น 3 ส่วนคือ ส่วนวิเคราะห์หาส่วนประกอบของเอกสาร ส่วนรู้จำตัวอักษรเดี่ยว และส่วนแยกตัวอักษรไทยที่ติดกัน โดยส่วนวิเคราะห์เอกสารจะเน้นที่การแก้ความเอียงซึ่งกำหนดให้เอียงได้ไม่เกิน 5 องศา เมื่อได้เอกสารภาพแล้วทำการแยกบรรทัดโดยใช้ โปรเจกชั่นแนวนอน และดิ่งเพื่อแยกตัวอักษรออกเป็นบรรทัด และคอลัมน์ตามลำดับ โดยในการแยกคอลัมน์นั้นใช้โปรเจกชั่นแนวดิ่ง(Vertical Projection) ในการหาจุดตัดแยกและอาศัยการรวมจุดในแนวดิ่ง บริเวณที่ไม่มีจุดภาพเป็นจุดที่ใช้ตัดคอลัมน์ ส่วนการแยกบรรทัดนั้นใช้โปรเจกชั่นแนวนอน(Horizontal Projection) ซึ่งใช้การนับการรวมจุดในแนวนอน โดยบริเวณที่ไม่มีจุดภาพใดๆ จะถือเป็นจุดที่ใช้ตัดคอลัมน์ไม่ได้ เนื่องจากในกรณีที่บรรทัดเอียงเล็กน้อยจะทำให้ไม่มีช่องทะลุที่จะใช้ตัดบรรทัดได้ ดังนั้นจึงต้องใช้ช่องระหว่างภูเขาที่เกิดจากการ Project แต่ปัญหาเกิดขึ้นอีกว่าใช้เกณฑ์อะไรบอกว่าจุดนั้นต่ำพอที่จะเป็นหุบเขาได้ ซึ่งทำได้ง่ายโดยเทียบกับยอดเขา แต่ปัญหาต่อมาเกิดอีกว่าถ้าบรรทัดมีอักษรน้อย ยอดเขาจะต่ำ อาจถูกมองว่าเป็นหุบเขาได้ เราจึงต้องปรับปรุงการทำ Horizontal Projection ในการทำ Horizontal Projection ตามปกตินั้นเราจะได้ลักษณะภูเขา 1 ลูกต่อ 1 บรรทัด ถ้าบรรทัดยาวใกล้เคียงกันความสูงของยอดภูเขาแต่ละลูกจะใกล้เคียงกัน แต่ในกรณีที่บางบรรทัดสั้น เช่น จบย่อหน้า หัวเรื่อง จะทำให้ภูเขาบางลูกเล็กจะทำการปรับ Horizontal Projection โดยเรียกว่า Modified Horizontal Projection ช่วงที่มีค่ามากที่สุดของแต่ละบรรทัด จากนั้นทำการหาจุดตัดโดยตรวจสอบหาหุบเขาซึ่งใช้หลักการว่า "จุดตัด = จุดที่ต่ำที่สุดระหว่างความชันที่เป็นลบ ซึ่งต่อเนื่องจากยอดถึงพื้นราบกับความชันที่เป็นบวกซึ่งต่อเนื่องจากพื้นราบถึงยอดเขา" เมื่อได้ตัวอักษรเดี่ยว หรือตัวอักษรที่ติด แล้วทำการหาระดับเพื่อแยกตัวอักษรออกเป็นกลุ่มก่อนทำการตัดแยกเป็นอักษรเดี่ยว โดยการตรวจสอบจะทำการหา ค่า topline และ baseline ซึ่งทั้งสองค่าได้จากการตัดยอดภูเขาที่ได้จาก Horizontal Project แล้วทำการตรวจสอบลักษณะตัวอักษรเพื่อแบ่งแยกออกเป็นกลุ่มๆ โดยพิจารณาหาจุดตัดในแต่ละกลุ่มก่อนที่จะส่งให้ส่วนรู้จำตัวอักษรทำการวิเคราะห์ออกมาเป็น ตัวอักษรเท็กซ์ จากการทดลองพบว่าที่ขนาดตัวอักษร 12 pt จะมีตัวอักษรที่ติดกัน 7.6% สามารถทำการตัดแยกจากเอกสารทั้งหมด(อักษรที่ติดและ ไม่ติดกัน)ได้ถูกต้อง 95.6% ตัวอักษรขนาด 14 pt มีตัวอักษรที่ติดกัน 4.8% สามารถทำการตัดแยกจากเอกสารทั้งหมดได้ 96.5%

กิตติพงษ์ เจนวิถีสุข [2] ได้ทำการศึกษารู้จำตัวอักษรภาษาไทยโดยใช้นิวรอลเน็ตเวิร์กและ วิธีซินแทกติก โดยแบ่งเป็น 3 ขั้นตอนคือ Preprocessing, Feature/Primitive Extraction และ Neural Network ในส่วนของ Preprocessing ได้เตรียมข้อมูลที่นำเข้าเป็นตัวอักษรที่ตัดเป็นตัวอักษรเดี่ยวๆ แล้วนำมากำจัดสัญญาณรบกวนโดยตรวจสอบจุดข้างเคียงรอบๆจุดที่ทำการตรวจสอบ จำนวน 8 จุด โดยถ้าไม่มีจุดดำรอบๆจุดดังกล่าวให้ถือเป็นสัญญาณรวบกวน จากนั้นทำตัวอักษรให้บาง (Thinning) โดยเปลี่ยนแปลงข้อมูลให้เหลือเฉพาะเส้นโครงร่างของภาพตัวอักษรก่อนทำการส่งไปยังส่วนการรู้จำ จากการทดสอบกับตัวอักษรภาษาไทยจำนวน 1,392 ตัวอักษร มีความถูกต้องเฉลี่ย 99.28%

สนธยา เมรินทร์ [4] ทำการรู้จำตัวอักษรภาษาไทยโดยใช้วิธีซินแทกติก ซึ่งเป็นการพิจารณาที่โครงสร้างของตัวอักษร โดยมีการอธิบายโครงสร้างของตัวอักษรในรูปของประโยคที่ประกอบด้วย primitive ทำให้สามารถจำแนกตัวอักษรที่มีโครงสร้างที่แตกต่างกันออกจากกันได้ และยังใช้วิธีการเปรียบเทียบทาง feature สำหรับกรณีตัวอักษรภาษาไทยบางกลุ่มที่มีลักษณะคล้ายคลึงกันมาก

งานวิจัยของ วัชระ ฉัตรวิริยะ [5] ทำการรู้จำตัวอักษรภาษาไทยเป็นบรรทัดโดยเน้นที่การตัดแยกอักษรที่ติดกัน มีการกำหนดขนาดตัวอักษร วิธีการทำงานของโปรแกรมคือ การแยกภาพตัวอักษรออกจากบรรทัดโดยดูช่องว่างในแนวนอน ดังนั้นจึงตรวจสอบว่าเป็นอักษรที่ติดหรือไม่โดยใช้จำนวนความกว้างจุดภาพสำหรับการตัดนั้นจากอักษรที่สงสัยว่าติดมาเทียบกับฐานข้อมูลที่สร้างเอาไว้ เมื่อตรงกับแบบที่ติดแบบใดแล้วจึงตัดตามแบบนั้น ฐานข้อมูลที่ว่านี้คือการหาโปรเจกชั่นของภาพตามแนวนอน และแนวตั้ง สำหรับตัวติดทุกประเภท อักษรที่เข้ามากับฐานข้อมูลนั้นใช้การเปรียบเทียบแบบจุดต่อจุด ทำให้น่าสงสัยว่าถ้ามีแบบอักษรที่ติดมากๆ จะใช้ได้ผลหรือไม่ เช่น "ณ" มีลักษณะของโปรเจกชั่นเหมือนกับตัว "เม"

งานวิจัยของ Nucharee Premchiwadi, Wichian Premchaiwadi และ Seinosuke Narita [7] ทำการตัดแยกตัวอักษรที่ติดกันโดยใช้วิธี Horizontal and Vertical Projection เริ่มด้วยการแบ่งตัวอักษรออกเป็นกลุ่มๆ แล้วกำหนดกลุ่มตัวอักษรที่สามารถติดกันได้ เมื่อได้กลุ่มตัวอักษรแล้วทำการตัดแยกแนวดิ่งโดยหาบริเวณที่มีจำนวนจุดรวมน้อยที่สุดเพื่อใช้เป็นจุดตัด โดยการตรวจสอบว่าเป็นตัวอักษรที่ติดกันหรือ ไม่สามารถตรวจสอบได้จากค่าความสูงของตัวอักษรต้องไม่เกิน 1.5 ของตัวอักษรในระดับบน ส่วนแนวนอนใช้วิธีของ Kahan and Pavlidis ซึ่งพัฒนาต่อจาก Lu โดยค่า PV คำนวณได้จากการหาค่าเฉลี่ยของจุดข้างเคียงก่อน เพื่อแก้ไขข้อผิด

พลาดจากกรณีที่มีรอยเปื้อนหรือ มีการติดกันมากๆ ก่อนทำการหาอนุพันธ์อันดับที่สองของค่าโปรเจกชั่นในแนวดิ่ง กับค่าโปรเจกชั่นในแนวดิ่ง โดยตัวอักษรที่นำมาพิจารณาในการตัดนั้นจะต้องมีความกว้างไม่ต่ำกว่า 0.8 ของความสูงของตัวอักษร