Analyse de manuscrits sur feuilles de palmier numérisés de l'Asie du Sud-Est : de la reconnaissance de glyphes à la translitération automatique

Made Windu Antara KESIMAN, Jean-Christophe BURIE, Jean-Marc OGIER, Philippe GRANGÉ

Laboratoire Informatique, Image et Interaction (L3i)
Université de La Rochelle, France

Mots-clés: *manuscrits sur feuilles de palmier, segmentation, reconnaissance de glyphes, translitération automatique, écriture balinaise, dataset*

I. INTRODUCTION

A. Contexte

Un patrimoine culturel très précieux, témoin de la richesse de la littérature ancienne d'Asie du Sud-Est est conservé dans des manuscrits rédigés sur des feuilles de palmier. La plupart des œuvres littéraires ont été écrites sur des feuilles de palmiers séchées et traitées (Fig. 1), avec des langues et des scripts complexes de l'Asie du Sud-Est. De nombreuses collections très importantes de manuscrits historiques ont été trouvées en Asie du Sud-Est, par exemple en Thaïlande [7], au Cambodge [5], et aussi en Indonésie [1]. A Bali, ces manuscrits (*Lontars*) recèlent des textes littéraires anciens composés dans la vieille langue javanaise Kawi avec de nombreux emprunts lexicaux au Sanskrit. Malheureusement, de nombreux *Lontars*, qu'ils fassent partie des collections de certains musées ou appartiennent à des familles privées, sont dans un état de dégradation inquiétant, à cause du temps et/ou des mauvaises conditions de conservation. Les matériaux naturels tels que les feuilles de palmier ne résistent pas à l'usure du temps, et donc les efforts de numérisation et d'ĭndexation de *Lontars* sont vitaux pour sauvegarder ce patrimoine. Au cours des cinq dernières années, les collections de manuscrits sur feuilles de palmier d'Asie du Sud-Est ont reçu une grande attention de la part des chercheurs dans le domaine de l'analyse d'images de documents. Pour préserver ces précieuses collections, certains projets de numérisation de ces manuscrits ont déjà été proposés. Mais pour ouvrir, au grand public, un accès plus large au contenu de ces manuscrits, un système de translitération pour convertir les scripts originaux en scripts Romain est nécessaire. Par exemple, dans le cas des manuscrits de Bali en Indonésie, le système de translitération sera très utile aux jeunes chercheurs qui ne connaissent pas le script balinais, pour lire ces manuscrits. Nos recherches abordent non seulement les problématiques de la numérisation des manuscrits sur feuilles de palmier, l'analyse automatique et la mise au point d'un système d'ĭndexation efficaces de ces manuscrits. En ajoutant, un système de reconnaissance de glyphes ou de mots ainsi qu'une méthode de translitération automatique pour convertir les anciens alphabets vers l'alphabet latin, le traitement de ces manuscrits permettrait de les valoriser, en facilitant la diffusion du contenu de ces œuvres. Bien que certaines méthodes d'analyse aient maintenant atteint une performance satisfaisante en particulier pour l'alphabet latin, la nécessité de développer des méthodes d'analyse de documents et d'indexation pour d'autres types de documents historiques reste un défi majeur pour les chercheurs dans le domaine de l'analyse des documents.

Fig. 1. Manuscrit sur feuille de palmier de Bali

B. Les objectifs

L'objectif principal de cette recherche est d'apporter une valeur ajoutée aux images numérisées des manuscrits sur feuilles de palmier en développant des outils pour analyser, indexer et fournir un accès rapide et efficace au

contenu des *Lontars*. Ainsi, les *Lontars* seraient plus accessibles, lisibles et compréhensibles pour un public plus large. Pour ouvrir, au plus grand nombre, l'accès du contenu précieux de ces manuscrits historiques, il faut un système approprié pour translitérer le script balinais vers le script Romain. Pour atteindre cet objectif, un système de reconnaissance de glyphes balinais est très important. Ce schéma doit être développé en tenant compte de l'état dégradé des manuscrits écrits sur des feuilles de palmier et de la complexité du script balinais. L'un des principaux problèmes pour l'analyse automatique et l'indexation des manuscrits réside dans les caractéristiques physiques et l'état des manuscrits. Généralement, les manuscrits sur feuilles de palmier présentent une mauvaise qualité de lecture, car les supports se sont dégradés au fil du temps en raison des conditions de stockage. Les images des manuscrits sur feuilles de palmier affichent des textes décolorés en raison du vieillissement et un faible contraste, très souvent avec des taches. De nombreuses déformations dans la forme des caractères apparaissent sur les images numérisées en raison de la proximité des caractères. Enfin l'espace entre les lettres et les lignes est très variable. Dans le cadre de cette recherche, nous nous concentrons sur ces caractéristiques qui constituent un défi pour l'évaluation de la robustesse des méthodes d'analyse de documents et d'indexation des manuscrits. Les méthodes classiques sont inefficaces sur les manuscrits sur feuilles de palmier. De nouvelles approches sont donc nécessaires pour indexer ces images. Nos travaux portent sur plusieurs tâches de traitement d'image, de la numérisation du document à la reconnaissance de caractères ou de mots (*word spotting*), en passant par la translitération automatique, et jusqu'à la création d'un moteur de recherche pour naviguer dans ces collections de manuscrits.

C. Travaux précédents

La numérisation, la transcription, et l'annotation des images de ces manuscrits ont été réalisées à Bali. Tous ces jeux de données ont été publiés après la compétition officielle que nous avons organisée à la conférence internationale ICFHR[1][1,2]. Nous avons proposé un modèle spécifique pour construire la vérité terrain des images binarisées et ainsi évaluer plus facilement les méthodes de binarisation développées pour les manuscrits sur feuilles de palmier [3]. Une étude importante a également été conduite sur l'intervention humaine et l'effet de la subjectivité sur la construction de la vérité terrain. Cette expérience, réalisée en condition réelle a permis d'analyser et mesurer quantitativement l'existence de la subjectivité humaine dans la construction de la vérité terrain des images binarisées [4]. Ces manuscrits constituent un défi pour la segmentation des lignes de texte. Nous avons étudié la performance de six méthodes de segmentation de lignes de texte en menant des études expérimentales comparatives sur notre collection d'images manuscrites. Le corpus d'image utilisé dans cette étude provient d'images issues de trois types de manuscrits d'Asie du Sud-Est: balinais, sundanais (manuscrits d'Indonésie) et khmer (Cambodge) [5]. Un système de reconnaissance de glyphes et de translitération automatique est essentiel pour valoriser ces collections d'images. Nous avons donc mené une étude expérimentale sur les méthodes d'extraction de caractéristiques pour la reconnaissance des glyphes du script balinaise [6]. Nous présentons dans cet article un schéma de la translitération automatique pour ces manuscrits.

D. Schéma général proposé

Compte tenu de l'état dégradé des manuscrits écrits sur les feuilles de palmier et de la complexité du script balinais, nous présentons dans cet article, un schéma complet pour la reconnaissance de glyphes basé sur la catégorisation spatiale et l'utilisation de règles phonologiques pour la translitération de ces manuscrits (Fig. 2). En partant de l'image des manuscrits, ce schéma est initialisé avec un processus de segmentation des lignes de texte et des glyphes. Puis il est suivi d'un processus de reconnaissance des glyphes. Dans ce schéma, une reconnaissance globale et cinq différentes méthodes de reconnaissance catégorisées de glyphes ont été combinés lors de l'étape de vérification et de validation pour produire trois options de reconnaissance de glyphes. Ces cinq différentes approches de reconnaissances de glyphes prennent en compte la position spatiale de chaque glyphe sur le manuscrit. Elles sont utilisées pour vérifier et valider le résultat de reconnaissance globale du glyphe. Chaque approche pour reconnaitre les glyphes est construite en combinant plusieurs méthodes d'extraction de caractéristiques et un réseau de neurones à une seule couche. Le réseau neuronal est initialisé par un apprentissage de caractéristiques non supervisé. Des règles sont appliquées pour sélectionner l'option de reconnaissance de glyphe la plus appropriée et sont également utilisées pour détecter l'erreur potentielle lors de la segmentation et la reconnaissance du glyphe. Enfin, le résultat du système de reconnaissance du glyphe qui est l'option sélectionnée,

[1] http://amadi.univ-lr.fr/ICFHR2016_Contest/

est envoyé au module de translitération phonologique, pour être translitérée en script Romain. Les résultats ont été évalués avec la vérité terrain du texte translitéré fournie par les philologues. Le schéma proposé fournit des résultats très prometteurs pour la translitération des manuscrits balinais et peut être adapté à d'autres types de script.

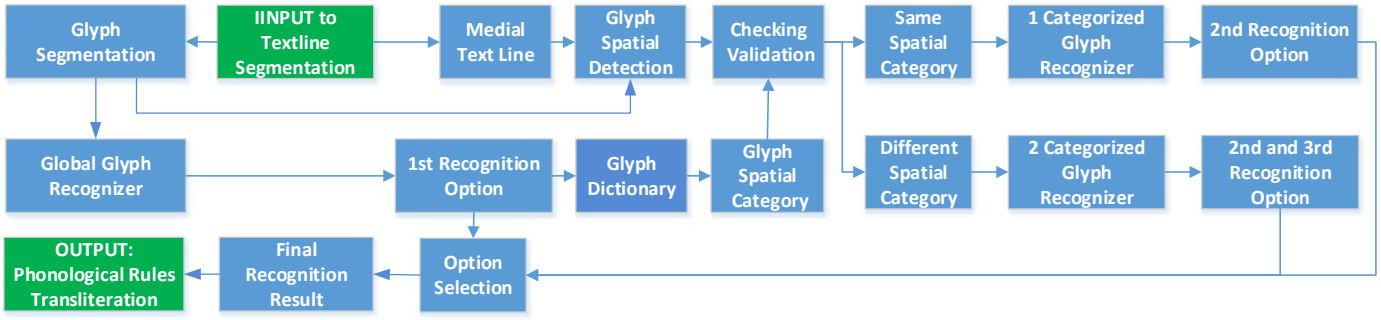

Fig. 2. Schéma proposé

Cet article est organisé ainsi : la section II présente brièvement les manuscrits écrits sur des feuilles de palmier et le script balinais. La section III détaille le schéma proposé pour la reconnaissance de glyphes et la translitération des manuscrits balinais. Le jeu de données AMADI_LontarSet [1] qui a été utilisé dans les expérimentations et l'évaluation des résultats expérimentaux est présenté dans la section IV. Enfin, une conclusion et quelques perspectives pour les futurs travaux sont données à la section V.

II. MANUSCRIT SUR DES FEUILLES DE PALMIER ET SCRIPT BALINAIS

A. Manuscrit sur des feuilles de palmier en Asie du Sud-Est

Dans les manuscrits Khmer du Cambodge et dans les manuscrits de Bali et de Sunda d'Indonésie, le système d'écriture est basé sur un alphabet syllabique. Cependant dans ces scripts, certains glyphes sont écrits au-dessus la ligne médiane de texte (type ASCENDER) ou sous la ligne médiale de texte (type DESCENDER). La position spatiale de chaque glyphe relativement à la ligne médiane de texte sur le manuscrit peut être utilisée de manière appropriée comme une information importante dans la reconnaissance du glyphe. Plus généralement, le schéma proposé peut être appliqué à ces types de script. Dans cet article, nous nous concentrons sur le script balinais.

B. Script balinais

À Bali, en Indonésie, les manuscrits de feuilles de palmier s'appelent *Lontar* (Fig. 1 & 3). Les *Lontars* ont été écrits en écriture balinaise et en langue balinaise, composées dans l'ancienne langue javanaise de Kawi et Sanskrit. La langue balinaise est une langue malayo-polynésienne parlée par plus de 3 millions de personnes principalement à Bali, en Indonésie. Le script balinais (*Aksara Bali*) est un descendant du script Brahmi, et est considéré comme l'un des scripts les plus complexes de l'Asie du Sud-Est. L'alphabet et les chiffres du script balinais sont composés de ±100 classes de caractères comprenant des consonnes, des voyelles et d'autres caractères composés spéciaux. Conformément à l'Unicode Standard 9.0, le script balinais a en fait la table Unicode de 1B00 à 1B7F.

III. SCHÉMA COMPLET POUR LA RECONNAISSANCE DE GLYPHES ET POUR LA TRANSLITÉRATION

Le schéma proposé se compose de six tâches: la segmentation de la ligne de texte et du glyphe, la détection de la position spatiale pour la catégorisation des glyphes, le processus d'ordonnancement des glyphes, la reconnaissance globale et catégorisée du glyphe, la sélection de l'option pour la reconnaissance du glyphe et la translitération basée sur des règles phonologiques.

A. Segmentation de la ligne de texte et du glyphe

Les performances de six méthodes de segmentation de lignes de texte sur les images de manuscrit sur des feuilles de palmier de l'Asie du Sud-Est ont déjà été étudiées par Kesiman et al [5]. Sur la base de ces études expérimentales comparatives, chaque méthode a été construite de manière optimale en se basant sur certaines caractéristiques spécifiques de la collection de manuscrits. Le comportement de certaines méthodes est fortement influencé par certains défis qui sont clairement présents sur chaque collection de manuscrits de l'Asie du Sud-Est. Pour la collection de manuscrits balinais, la méthode de *Seam Carving* [8–10] fournit le meilleur résultat sur la segmentation de ligne de texte en niveaux de gris [5].

La méthode de *Seam Carving* détermine le chemin de segmentation en fonction d'une *Seam Map* définie qui est générée à partir d'une fonction d'énergie. Dans cette méthode, deux types de *Seam* sont calculés: les *Seams* médianes et les *Seams* séparées (Fig. 3). Dans notre schéma, l'implémentation de la méthode *Seam Carving* [8] a été appliquée non seulement pour segmenter les lignes de texte, mais aussi pour segmenter les zones de glyphe sur chaque ligne de texte (Fig. 3).

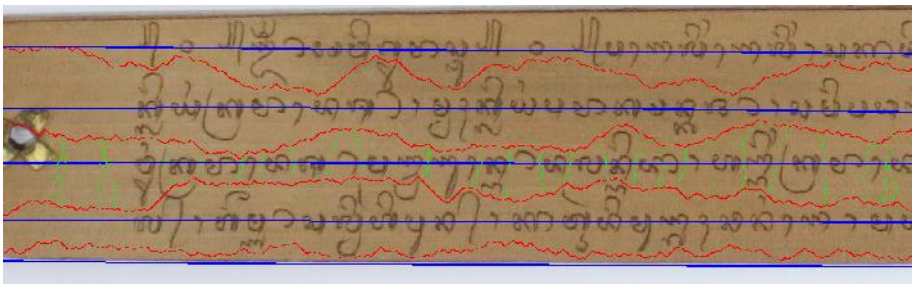

Fig. 3. Segmentation de lignes de texte avec la méthode de *Seam Carving* : *Seams* médianes (bleu), *Seams* séparées (rouge) et zones de glyphes (en vert sur la 3eme ligne).

Pour segmenter les zones de glyphe, la méthode du *Seam Carving* s'applique verticalement sur chaque ligne de texte. Il faut noter qu'une zone de glyphe peut contenir plus d'un glyphe. La binarisation avec la méthode d'Otsu est ensuite effectuée sur cette zone locale de l'image du manuscrit original en niveaux de gris. L'analyse des composantes connexes est finalement utilisée pour nettoyer et extraire tous les segments de glyphe dans chaque zone de glyphe (Fig. 4). Un processus de nettoyage est appliqué pour supprimer certains petits bruits du processus de binarisation. Une composante connexe est considérée comme un bruit si la taille (hauteur ou largeur) est inférieure à un seuil défini comme taille de glyphe minimum possible dans le jeu de donnée AMADI_LontarSet [1] (voir la section IV.A).

B. Détection de la position spatiale pour la catégorisation de glyphes

Pour chaque glyphe détecté dans une zone de glyphe, leur position spatiale est définie par rapport à la ligne médiane de texte sur le manuscrit. Six catégories de position spatiale ont été définies (Fig. 5). Les glyphes *Ascender* (ASC) et les glyphes *Descender* (DESC) ne croisent pas la ligne médiane de texte. Les glyphes ASC sont écrits au-dessus de la ligne médiane de texte, et les glyphes DESC sont écrits sous la ligne médiane. Les glyphes BASE sont normalement les glyphes de base écrits exactement sur la ligne médiane, tandis que les glyphes ASC-BASE (respectivement BASE-DESC) sont les glyphes composés d'un glyphe BASE et d'un glyphe ASC (respectivement DESC). Dans notre collection de glyphes, il n'y a qu'une classe de glyphe ASC-BASE-DESC

appelée glyphe "ADEG-ADEG". Ces catégories fonction de la position spatiale sont utilisées pour construire le système de reconnaissance de glyphe par catégorie et pour structurer les règles phonologiques dans le processus de translitération.

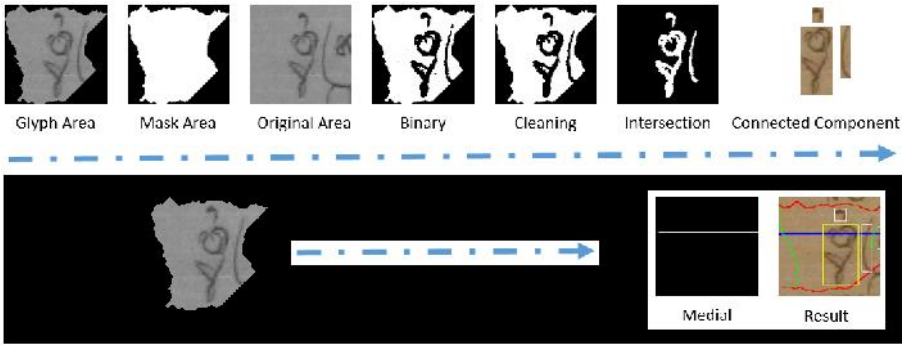

Fig. 4. Détection de la zone de glyphe et le processus de la segmentation de glyphe

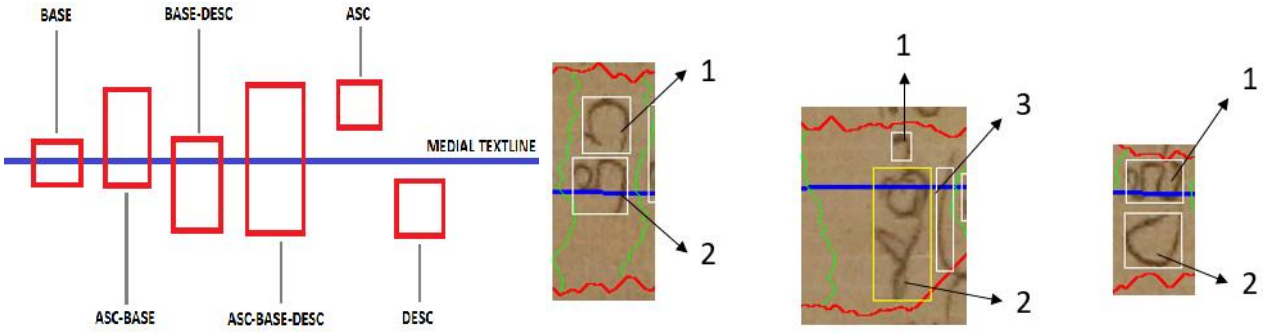

Fig. 5. Gauche: le position spatiale des glyphes est définie par rapport à la ligne médiane de texte. Droite: Exemples de d'ordonnancement des glyphes.

C. Processus d'ordonnancement des glyphes

Dans le script balinais, il n'y a pas d'espace entre les mots. Une fois que les glyphes ont été segmentés dans chaque zone de glyphe, le processus d'ordonnancement des glyphes est alors appliqué. L'ordre des glyphes est très important pour l'utilisation des règles phonologiques (voir la section III.F) lors de la phase de translitération du script balinais.

Les glyphes qui se trouvent sur la ligne médiane de texte (BASE ou ASC-BASE ou BASE-DESC ou glyphe ASC-BASE-DESC, voir section III.B) sont positionnés de gauche à droite en fonction de leur position par rapport à la bordure gauche de la zone de glyphe. S'il y a un glyphe ASC et/ou un glyphe DESC, le glyphe ASC sera placé avant le glyphe BASE associé, et le glyphe DESC sera placé après leur glyphe BASE associé. Cette règle de positionnement est appelée simplement « ordre BASE-ASC-BASE-DESC-BASE". En raison des différents styles d'écriture de manuscrit, la position et la taille des glyphes ASC ou DESC ne sont parfois pas exactement sur l'axe vertical par rapport à la position du glyphe BASE associée. Il est alors difficile de définir quel glyphe BASE appartient à quel glyphe ASC ou DESC.

Pour surmonter ce problème, la relation spatiale entre chaque couple de glyphe est définie en fonction de ses quatre positions par rapport à la bordure (supérieure, inférieure, gauche et droite). Deux glyphes peuvent donc avoir un lien vertical, horizontal ou diagonal. Le glyphe BASE le plus proche qui peut être associé à un ASC ou un glyphe DESC est ensuite sélectionné. Dans la Fig. 5, sur le premier exemple, le glyphe 1 (ASC) et le glyphe 2 (BASE) ont un lien vertical entre eux. Le glyphe 1 est lié au glyphe 2. Sur le deuxième exemple, le glyphe 1 (ASC) et le glyphe 2 (BASE-DESC) ont un lien vertical, le glyphe 1 et le glyphe 3 (BASE-DESC) ont un lien diagonal, et le glyphe 2

et le glyphe 3 ont un lien horizontal entre eux. Le glyphe 1 est lié au glyphe 2, mais le glyphe 1 n'est pas lié au glyphe 3. Alors que sur le troisième exemple, le glyphe 1 (BASE) et le glyphe 2 (DESC) ont un lien vertical entre eux. Le glyphe 2 est lié au glyphe 1.

D. Reconnaissance globale et catégorisation du glyphe

La performance de la reconnaissance de glyphe isolée dépend grandement de l'étape d'extraction des caractéristiques . À notre connaissance, seuls quelques systèmes sont disponibles dans la littérature pour la reconnaissance des scripts d'Asie du Sud-Est. Pour le script balinais, une première étude a été réalisée par Kesiman et al [6]. Ils ont évalué la performance individuelle de dix méthodes d'extraction de caractéristiques et ont proposé une combinaison appropriée et robuste de ces méthodes pour augmenter le taux de reconnaissance. Ils ont constaté que les méthodes *Histogramme de gradient orienté* (HoG) [11], *Neighborhood Pixels Weights* (NPW) [12], Contours directionnels de *Kirsch* [12] et la méthode du *Zoning* [12–14] donnent des résultats très prometteurs. Leur étude a montré que le taux de reconnaissance peut être considérablement augmenté en appliquant NPW sur quatre images directionnelles de *Kirsch*. De plus l'utilisation de NPW sur *Kirsch* en combinaison avec les méthodes de *HoG* et du *Zoning* peut augmenter le taux de reconnaissance jusqu'à 85.16% avec le classificateur des k plus proches voisins (k-NN).

Dans ce travail, la même combinaison de méthodes d'extractions de caractéristiques a été adoptée pour construire un système de reconnaissance de glyphes. Ce système est bâti sur un réseau neuronal à une seule couche et un apprentissage initial non supervisé basé sur la méthode des K-moyennes (Fig. 6). Ce schéma s'inspire de l'étude de Coates et al [15,16]. L'apprentissage non supervisé calcule les poids initiaux lors de la phase d'apprentissage du réseau neuronal à partir des centres de cluster de tous les vecteurs de caractéristiques.

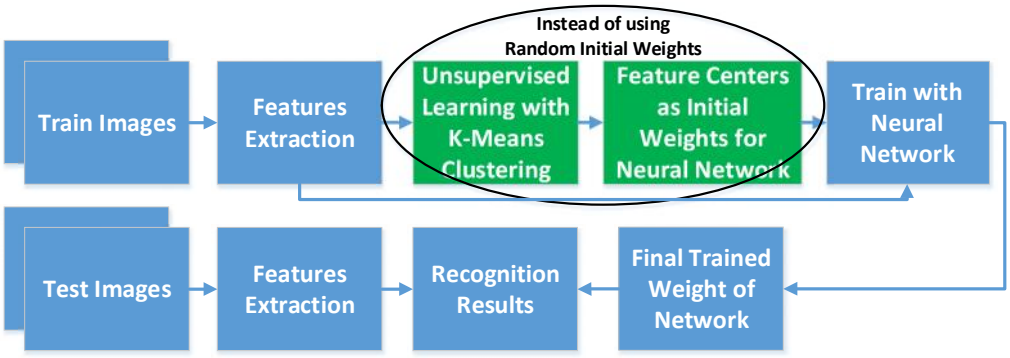

Fig. 6. Schéma du système de reconnaissance de glyphes

En utilisant ce schéma de reconnaissance de glyphe, un système de reconnaissance global et cinq systèmes de reconnaissance par catégorie de glyphes ont été construits. Le système global a été entrainé sur un total de 133 classes de glyphes de script balinais à partir du jeu de données AMADI_LontarSet [1] (voir la section IV.A). Les cinq systèmes de reconnaissance par catégorie de glyphes ont été entrainés uniquement sur des sous-ensembles de glyphes. Chaque sous ensemble correspond à une catégorie et regroupe les glyphes en fonction de leur position spatiale. Il faut noter que le nombre d'images d'échantillons pour chaque classe est différent. En effet, en langue balinaise, certaines classes sont fréquemment utilisées, mais d'autres le sont plus rarement.

E. Sélection de l'option pour la reconnaissance du glyphe

Le schéma de reconnaissance pour chaque glyphe segmenté est comme suit (voir Fig. 2, section I): envoyer le glyphe au système de reconnaissance global (considérer le résultat comme la 1ère option). Sur la base de ce premier résultat de reconnaissance globale, chercher dans le dictionnaire de glyphe pour savoir à quelle catégorie spatiale ce glyphe devrait normalement appartenir. Faire la deuxième reconnaissance en envoyant ce segment de glyphe au système de reconnaissance par catégorie du glyphe associé (considérer le résultat comme la 2ème option). Définir la catégorie spatiale réelle de cette position du segment de glyphe par rapport à la ligne médiane de texte. Cette catégorie spatiale réelle peut être identique ou différente avec la catégorie spatiale vérifiée dans le dictionnaire de glyphe. Faire la troisième reconnaissance en envoyant ce segment de glyphe au système de reconnaissance de glyphe par catégorie (considérer le résultat comme la 3ème option). Finalement, trois options de reconnaissance ont été définies pour chaque glyphe : reconnaissance globale (G), reconnaissance catégorisée basée sur le dictionnaire de glyphe (D) et reconnaissance catégorisée basée sur la position spatiale de glyphe (S). Les règles de sélection des options sont les suivantes:

Si la détection de la catégorie spatiale est la même avec le dictionnaire glyphe, il existe deux possibilités:

1) Si G = D = S, il n'y a qu'une seule option. La confiance dans la segmentation est forte et la reconnaissance considérée comme correcte. Le résultat de reconnaissance finale est G / D / S (Fig. 7a).

2) Si G < > (D = S), il existe deux options différentes. Les résultats de la reconnaissance sont différents entre le système global et le système par catégorie. Dans ce cas, le résultat de la reconnaissance finale est D / S.

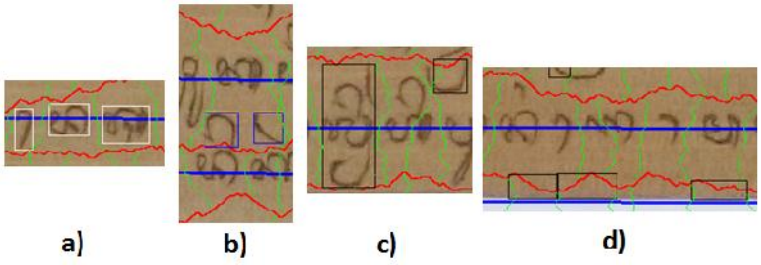

Fig. 7. a) Forte confiance dans la segmentation et la reconnaissance (blanc), b) Reconnaissance correcte mais avec une erreur potentielle dans la segmentation (bleu), c) & d) Mauvaise segmentation (noir)

Si la détection de la catégorie spatiale est différente avec le dictionnaire de glyphes, il existe trois possibilités:

1) Si (G = S) < > D, il existe deux options différentes. Ce cas est impossible car pour avoir G = S, G et D devraient avoir la même catégorie spatiale.

2) Si (G = D) < > S, il existe deux options différentes. Il existe en fait trois sous-cas à étudier. Si (S = BASE / ASC-BASE / BASE-DESC / ASC-BASE-DESC et D = ASC / DESC) ou vice versa, cela signifie qu'il existe une grande différence entre la détection de catégorie spatiale et le dictionnaire de glyphe. Le résultat de la reconnaissance finale est S. Si (S = ASC et D = DESC) ou vice versa, la reconnaissance est peut-être correcte, mais une mauvaise affectation de ligne de texte peut être détectée pour ce segment de glyphe (Fig. 7b). Le résultat de la reconnaissance finale est G / D. Pour tous les autres sous-cas, il peut s'agir d'une mauvaise segmentation du glyphe (Fig. 7c). Le résultat de la reconnaissance finale est G / D.

3) Si G < > D < > S, il existe trois options différentes. Cela implique que la ligne de texte ou la segmentation de glyphe sont erronées, et par conséquent la détection de la catégorie spatiale n'est pas correcte (Fig. 7d). Le résultat de la reconnaissance finale est G. Dans ce cas, nous pouvons envisager de relancer localement un traitement spécifique pour ce segment de glyphe.

Avant d'appliquer les règles de sélection des options finales, deux cas spéciaux sont traités. En effet, le glyphe "*CECEK*" et le glyphe "*TALENG*" peuvent être écrits dans deux différentes catégories spatiales. Pour ces deux glyphes, nous considérons seulement la catégorie spatiale dans le dictionnaire de glyphe.

F. Translitération basée sur des règles phonologiques

Pour compléter notre schéma, nous proposons une implémentation de la représentation des connaissances et des règles phonologiques pour la translitération automatique des manuscrits balinais sur feuilles de palmier. Les modules d'ordonnancement des glyphes, de reconnaissance de glyphes et de sélection des options de glyphe (voir la section III.C-E) alimentent le processus de translitération avec la structure de données de glyphe séquentielle présentée à la Fig. En balinais, le son final pour la syllabe du glyphe courant (CURR) de base (BASE) sera déterminé par l'*Ascender* (ASC) du glyphe courant, *Descender* (DESC) du glyphe courant, le BASE du glyphe suivant (NEXT), BASE du glyphe précédent (PREV), ou même dans certaines règles phonologiques, peut également être influencé par la BASE des deux glyphes précédents (PREV2) (Fig. 8). Les règles phonologiques sont finalement construites et définies formellement en fonction de cette structure de données construite à partir des résultats de reconnaissance de glyphes.

	1	2	3	4
1	[]	[]	'ULU'	[]
2	'SURANG'	'DI'	'RA'	'CECEK'
3	[]	[]	[]	[]

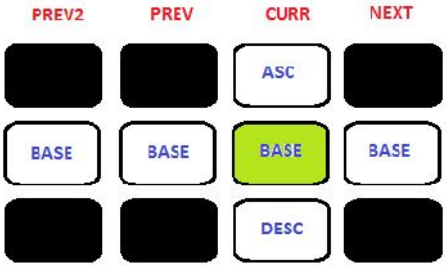

Fig. 8. Structure de données à partir des résultats de reconnaissance de glyphes pour la translitération

Un système basé sur les règles phonologiques pour effectuer des translitérations est proposé. Notre modèle est construit sur la phonétique qui est basée sur l'étude linguistique traditionnelle de la translitération balinaise. A partir de l'étude des segments de glyphe extraits de notre jeu de données AMADI_LontarSet [1] (voir la section IV.A), les propriétés du glyphe (composante de la syllabe: début (ONSET), noyau (NUCLEUS) et coda), les catégorisations de glyphe (consonnes, voyelles, conjonction, numéral, ponctuation et six catégories de position spatiale) dans le dictionnaire de glyphe[2], le schéma conditionnel des règles phonologiques pour la translitération du script balinais a finalement été identifié et formellement défini. Un exemple de règle phonologique est donné ci-dessous[3]:

```
RULE8:  IF  PREV.BASE.LEVEL1~="TALENG"  AND  CURR.ASC.LEVEL1=EMPTY  AND
CURR.BASE.LEVEL1~=EMPTY  AND  CURR.BASE.LEVEL2=CON  AND  CURR.BASE.LEVEL3=BASE  AND
CURR.BASE.ENDSYLLABLE="A" AND CURR.DESC.LEVEL1=EMPTY AND NEXT.BASE.LEVEL1~="ADEG-ADEG"
AND NEXT.BASE.LEVEL2~=GEM => SPEECH_SOUND=SPEECH_SOUND+"A"
```

Les règles phonologiques définissant la translitération sont fournies et sont appliquées dans un ordre séquentiel. Le système vérifie toutes les règles phonologiques conditionnelles qui pourraient être appliquées pour une position de glyphe donnée et les applique séquentiellement un par un.

IV. JEU DE DONNÉES ET RÉSULTATS EXPÉRIMENTAUX

A. Jeu de donnnées :AMADI_LontarSet

Dans nos expérimentations, nous avons utilisé AMADI_LontarSet [1], le premier jeu de données pour les manuscrits balinais sur des feuilles de palmier. Ce jeu de données est construit à partir de cent pages de manuscrit, sélectionnées issues de différentes collections de manuscrits de Bali. Ce jeu de données est publiquement

[2] Contact the authors to get the complete XML file of glyph dictionary

[3] Contact the authors to get the complete phonological rules

disponible pour la recherche scientifique[4]. Ce jeu de données annoté de glyphes isolés est constitué de 133 classes. Il comprend 11710 images pour le jeu d'apprentissage et 7673 images pour le jeu de test. Il a été utilisé pour les expérimentations de reconnaissance globales et catégorisées de glyphes isolés (voir la section III.D). 19 pages de manuscrit avec une annotation complète des segments de glyphe ont été utilisées pour tester et évaluer notre schéma de segmentation et de reconnaissance de glyphes. Enfin, 390 pages de manuscrits provenant de 23 collections différentes avec 1266 lignes de texte ont été utilisées pour tester et évaluer le schéma proposé pour la translitération. Pour ce jeu de données, la vérité terrain du texte translitéré a été effectuée par un expert : un philologue balinais.

B. Méthodes d'évaluation

1) Reconnaissance du glyphe isolé

Le taux de reconnaissance est calculé. Il est défini comme le pourcentage d'échantillons correctement reconnus sur les échantillons du jeu de données de test.

2) Schéma de segmentation et de reconnaissance de glyphes pour les pages de manuscrit

Le taux de segmentation (SR=*Segmentation Rate*) est calculé. Il est défini comme le pourcentage de segments de glyphe correct entre le fichier de résultat et le fichier de vérité de terrain (NSO=*Number Segment Overlapped*) sur le nombre total de segments de glyphe dans le fichier de résultat (NSR=*Number Segment Result*). La segmentation est considérée comme correcte si le taux de chevauchement entre la boite englobante du résultat et celle de la vérité terrain est supérieur à 50 %. Nous ne pouvons pas calculer le SR sur le nombre total de segments de glyphe dans le fichier de vérité terrain (NSG=*Number Segment Groundtruth*), car dans le fichier de vérité terrain, il peut y avoir deux segments de glyphe chevauchant une même zone. Le taux de reconnaissance segmenté (SRR=*Segmented Recognition Rate*) est ensuite calculé. Il est défini comme le pourcentage de segments de glyphe correctement reconnus dans le fichier de résultat (*NRR*) sur le *NSO* (Fig. 9).

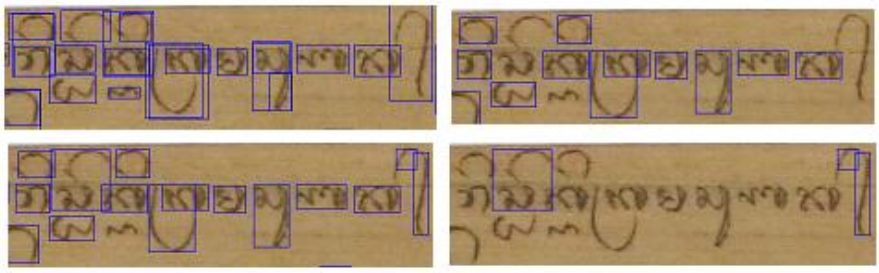

Fig. 9. Haut-gauche: segments de glyphe de vérité de terrain, Bas-gauche: segments de glyphe résultats, Haut-droite: segments de glyphe avec un chevauchement correct, Bas-droite: mauvais segments de glyphe

3) Translitération du manuscrit

Le *rappel du taux de motifs* (RPR) et la *précision du taux de motif* (PPR) sont calculés pour chaque ligne de texte dans la page de manuscrit. Le *taux de motif* est défini comme le pourcentage du même motif de texte entre le texte de résultat de translitération et le texte translitéré de vérité terrain, sur la longueur du texte translitéré de vérité terrain pour RPR ou sur la longueur du résultat de translitération pour PPR (Fig. 10).

GT:73.5484
tadharMANINGATUNGGATUNGga,leGASARATADANASARWAbARANAarTAMASwastramulIADAWALA.
yakaTAMADIANINGASARAPATUNGAN.NISTASARATUNGA,legADANASARWAboJARAPAna.uTAMANINgsa
Result:56.3758
wadhewrene0,wa**NGATUNGA**,2**GASARA**s**ADANA**wwabh**ARANA**h**ARWAMAS**umul**ADAWALA**,,
yenad**AMADhIANIASARAPA**wa**NGANINi**,jah,**NISTASARATU**2gnabhanayarwuane**JARANA**,,3**TAMANI**sa4a

[4] http://amadi.univ-lr.fr/ICFHR2016_Contest/index.php/download-123

Fig. 10. Motifs de texte (en rouge) extraits entre le texte de résultat de la translitération et le texte translitéré de la vérité terrain

Le motif de texte est extrait en générant l'arbre de suffixe généralisé [17] entre deux textes translitérés. La longueur minimale du motif de texte est de quatre caractères. Il est défini comme la longueur minimale d'un mot translitéré possible en balinais qui se compose de deux glyphes de base.

C. Résultats

1) Expérimentation sur la reconnaissance de glyphe isolé

Tous les systèmes de reconnaissance de glyphes isolés ont été construits en considérant 500 centres de clusters lors de l'apprentissage non supervisé des caractéristiques et un réseau de neurones à une seule couche avec 500 neurones. Les taux de reconnaissance de chaque système de reconnaissance sont présentés dans les tableaux I et II. Il montre que le réseau neuronal à une seule couche, initialisé avec l'apprentissage non supervisé des caractéristiques (NN=*Neural Network* + UFL =*Unsupervised Feature Learning*) a généralement amélioré le taux de reconnaissance par rapport à la dernière étude de la reconnaissance de glyphe balinais avec le réseau neuronal convolutif (CNN=*Convolutional Neural Network*) et les k plus proches voisins (k-NN=*k-Nearest Neighbours*) [6]. Les taux de reconnaissance des systèmes de reconnaissance de glyphes par catégorie sont généralement plus élevés que le taux de reconnaissance du système global. Cela signifie que si la détection de la catégorie spatiale d'un glyphe est correcte, la reconnaissance de glyphe par catégorie reconnaîtra mieux le glyphe que le système global.

Table I. Taux de reconnaissance du système de reconnaissance global de glyphes

Glyph Classes	Number of Data	CNN [7]	k-NN [7]	NN	NN UFL
133	Train : 11,710, Test : 7,673	84.31	85.16	85.51	85.63

Table II. Taux de reconnaissance du système de reconnaissance par catégorie de glyphe

Category	Glyph Classes	Number of Subset Data	NN	NN UFL
ASC	7	Train : 860, Test : 921	92.73	93.16
DESC	20	Train : 1,860, Test : 593	85.84	88.03
BASE	49	Train : 5,070, Test : 4,392	87.46	87.43
ASC-BASE	16	Train : 1,170, Test : 208	75.48	75.96
BASE-DESC	40	Train : 2,550, Test : 1,309	86.40	86.63

2) Expérimentation sur le schéma de segmentation et de reconnaissance du glyphe pour les pages manuscrites

Le tableau III montre la valeur NSG, NSR, NSO, SR, NRR et SRR pour le jeu de test. Le SR et le SRR des manuscrits numérotés de 3 à 4 et de 9 à 16 sont suffisamment élevés. Ces manuscrits proviennent de la collection du musée et présente les meilleures conditions de préservation. Les autres manuscrits sont issus de collections familiales privées et présentent plus de dégradation. Il est donc plus difficile de segmenter et de reconnaître les glyphes. Pour les manuscrits numérotés de 10 à11 et de 17 à 19, ces manuscrits ont été écrits dans un format de colonne particulier et la segmentation des lignes de texte n'est pas réalisée de manière optimale, donc SR et SRR sont plus faibles. Une analyse similaire peut être faite pour les résultats de la translitération de manuscrit.

Le tableau IV présente les valeurs maximales, minimales et moyennes de RPR et PPR pour chaque collection de manuscrits à partir de jeu de test. Les collections numérotées de 2 à 11, 19 et de 21 à 22 correspondent aux collections du musée. Le RPR et le PPR de ces collections sont suffisamment élevés. Comme expliqué lors de l'expérimentation précédente, dans les collections numérotées 20 et 23, plusieurs pages ont été écrites dans un format de colonne particulier, si bien que le RPR et le PPR sont très bas. En effet sur ces pages, la tâche de segmentation des lignes de texte n'a pas fonctionné de manière optimale. Un cas particulier que nous avons également observé sur la collection numéro 17 est qu'elle contient plus de graphiques et la bordure de la page est très endommagée ce qui a perturbé le processus de numérisation. Dans ce cas particulier, la tâche de segmentation a échoué ce qui explique les mauvais résultats.

Table III. Résultats de la segmentation et de la reconnaissance de glyphes

No	Manuscript Page	NSG	NSR	NSO	SR	NRR	SRR
1	Bangli-P41	357	216	107	49,54	62	57,94
2	Bangli-P47	326	134	79	58,96	44	55,70
3	IIA-10-1534-P8	654	517	400	77,37	336	84,00
4	IIIC-24-1641-P8	699	502	388	77,29	266	68,56
5	JG-01-P3	653	544	383	70,40	296	77,28
6	JG-02-P6	148	130	40	30,77	11	27,50
7	JG-02-P7	160	140	57	40,71	30	52,63
8	JG-05-P8	600	405	246	60,74	164	66,67
9	MB-AdiParwa(Purana)-5338.2-IV.a-P30	587	423	329	77,78	256	77,81
10	MB-AjiGriguh-5783-107.2-P11	122	180	54	30,00	40	74,07
11	MB-AjiGriguh-5783-107.2-P8	175	142	31	21,83	18	58,06
12	MB-ArjunaWiwaha-GrantangBasaII-P15	302	294	191	64,97	143	74,87
13	MB-ArjunaWiwaha-GrantangBasaII-P28	310	223	169	75,78	121	71,60
14	MB-TaruPramana-P3	614	405	288	71,11	186	64,58
15	MB-TaruPramana-P4	340	426	187	43,90	131	70,05
16	MB-TaruPramana-P6	281	310	138	44,52	75	54,35
17	WN-P5b	369	317	146	46,06	65	44,52
18	WN-P7a	186	179	20	11,17	3	15,00
19	WN-P9a	191	279	61	21,86	26	42,62

Table IV. Résultats de la translitération de manuscrits

No	Manuscript Collection	Nb Textline	RPR			PPR		
			Min	Max	Avg	Min	Max	Avg
1	Bangli	270	0	69,23	6,26	0	53,13	7,36
2	IIA-10-1534	30	0	73,13	47,78	0	68,16	43,73
3	IIA-5-789	30	0	76,25	50,27	0	77,85	48,67
4	IIB-2-180	29	0	72,63	54,04	0	69,61	52,69
5	IIIB-12-306	30	0	67,78	34,84	0	64,91	33,43
6	IIIB-42-1526	30	0	78,61	51,50	0	71,43	50,06
7	IIIB-45-2296	29	0	65,09	40,29	0	65,61	39,15
8	IIIC-19-1293	30	0	60,38	35,52	0	63,40	36,56
9	IIIC-20-1397	29	0	75,58	34,44	0	60,48	34,46
10	IIIC-23-1506	18	0	54,46	27,98	0	56,82	25,02
11	IIIC-24-1641	26	0	60,44	40,61	0	54,49	41,91
12	JG-01	62	0	56,47	23,88	0	57,72	20,70
13	JG-02	22	0	11,11	0,80	0	22,86	1,53
14	JG-03	56	0	25,83	4,91	0	23,88	7,34
15	JG-04	35	0	14,49	1,37	0	15,38	1,64
16	JG-05	17	0	27,37	4,55	0	41,94	8,85
17	JG-06	6	0	0,00	0,00	0	0,00	0,00
18	JG-07	33	0	4,94	0,15	0	12,90	0,39
19	MB-AdiParwa(Purana)-5338.2-IV.a	157	0	83,13	37,43	0	72,44	36,54
20	MB-AjiGriguh-5783-107.2	17	0	4,71	0,28	0	10,81	0,64
21	MB-ArjunaWiwaha-GrantangBasaII	86	0	65,81	26,83	0	68,18	32,89
22	MB-TaruPramana	149	0	70,15	17,69	0	66,67	18,99
23	WN	75	0	20,00	0,72	0	10,53	0,68
	All: 390 pages	1266						

V. CONCLUSIONS ET FUTURS TRAVAUX

Un schéma complet d'analyse pour la reconnaissance de glyphes et les règles phonologiques pour la translitération des manuscrits sur feuilles de palmier est présenté dans cet article. Le schéma proposé se compose de six tâches: la segmentation des lignes de texte et des glyphes, la détection de la position spatiale pour la catégorisation des glyphes, le processus d'ordonnancement des glyphes, la reconnaissance globale et par catégorie des glyphes, la sélection de l'option pour la reconnaissance du glyphe et la translitération basée sur des règles phonologiques. Ce schéma montre des résultats très prometteurs pour la translitération des manuscrits balinais sur feuilles de palmier. Le schéma proposé est modulaire et peut donc être adapté à un d'autre type de script en adaptant les modules en fonction des caractéristiques du script étudié.

Remerciements

The authors would like to thank Museum Gedong Kertya, Museum Bali, and all families in Bali, Indonesia, for providing us the samples of palm leaf manuscripts, and the students from the Department of Informatics Education and the Department of Balinese Literature, Ganesha University of Education for helping us in ground truthing process for this research project. This work is also supported by the DIKTI BPPLN Indonesian Scholarship Program and the STIC Asia Program implemented by the French Ministry of Foreign Affairs and International Development (MAEDI).

Bibliographie

[1] M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, G.N.M.A. Wibawantara, I.M.G. Sunarya, AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset, in: 15th Int. Conf. Front. Handwrit. Recognit. 2016, Shenzhen, China, 2016: pp. 168–172. doi:10.1109/ICFHR.2016.39.

[2] J.-C. Burie, M. Coustaty, S. Hadi, M.W.A. Kesiman, J.-M. Ogier, E. Paulus, K. Sok, I.M.G. Sunarya, D. Valy, ICFHR 2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts, in: 15th Int. Conf. Front. Handwrit. Recognit. 2016, Shenzhen, China, 2016: pp. 596–601. doi:10.1109/ICFHR.2016.107.

[3] M.W.A. Kesiman, S. Prum, J.-C. Burie, J.-M. Ogier, An Initial Study On The Construction Of Ground Truth Binarized Images Of Ancient Palm Leaf Manuscripts, in: 13th Int. Conf. Doc. Anal. Recognit. ICDAR, Nancy, France, 2015.

[4] M.W.A. Kesiman, S. Prum, I.M.G. Sunarya, J.-C. Burie, J.-M. Ogier, An Analysis of Ground Truth Binarized Image Variability of Palm Leaf Manuscripts, in: 5th Int. Conf. Image Process. Theory Tools Appl. IPTA 2015, Orleans, France, 2015: pp. 229–233.

[5] M.W.A. Kesiman, D. Valy, J.-C. Burie, E. Paulus, I.M.G. Sunarya, S. Hadi, K.H. Sok, J.-M. Ogier, Southeast Asian palm leaf manuscript images: a review of handwritten text line segmentation methods and new challenges, J. Electron. Imaging. 26 (2016) 11011. doi:10.1117/1.JEI.26.1.011011.

[6] M.W.A. Kesiman, S. Prum, J.-C. Burie, J.-M. Ogier, Study on Feature Extraction Methods for Character Recognition of Balinese Script on Palm Leaf Manuscript Images, in: 23rd Int. Conf. Pattern Recognit., Cancun, Mexico, 2016.

[7] R. Chamchong, C.C. Fung, Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand, in: IEEE, 2012: pp. 588–593. doi:10.1109/ICFHR.2012.280.

[8] N. Arvanitopoulos, S. Susstrunk, Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts, in: IEEE, 2014: pp. 726–731. doi:10.1109/ICFHR.2014.127.

[9] R. Saabni, J. El-Sana, Language-Independent Text Lines Extraction Using Seam Carving, in: IEEE, 2011: pp. 563–568. doi:10.1109/ICDAR.2011.119.

[10] C. Stoll, Y. Xiao, Z.-H. Duan, Text Line Extraction Using Seam Carving, in: Int. Conf. Image Process. Comput. Vis. Pattern Recognit. 2015, n.d.

[11] A. Aggarwal, K. Singh, K. Singh, Use of Gradient Technique for Extracting Features from Handwritten Gurmukhi Characters and Numerals, Procedia Comput. Sci. 46 (2015) 1716–1723. doi:10.1016/j.procs.2015.02.116.

[12] S. Kumar, Neighborhood Pixels Weights-A New Feature Extractor, Int. J. Comput. Theory Eng. (2009) 69–77. doi:10.7763/IJCTE.2010.V2.119.

[13] M. Blumenstein, B. Verma, H. Basli, A novel feature extraction technique for the recognition of segmented handwritten characters, in: IEEE Comput. Soc, 2003: pp. 137–141. doi:10.1109/ICDAR.2003.1227647.

[14] M. Bokser, Omnidocument technologies, Proc. IEEE. 80 (1992) 1066–1078. doi:10.1109/5.156470.

[15] A. Coates, H. Lee, A.Y. Ng, An Analysis of Single-Layer Networks in Unsupervised Feature Learning, in: Proc. 14th Int. Conf. Artif. Intell. Stat. AISTATS, Fort Lauderdale, FL, USA, n.d.

[16] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D.J. Wu, A.Y. Ng, Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning, in: IEEE, 2011: pp. 440–445. doi:10.1109/ICDAR.2011.95.

[17] M.W.A. Kesiman, Extraction des chaînes des symboles dans les séquences, Master Thesis, Universite de La Rochelle, 2006.