

Reconnaissance de chiffres farsi isolés par réseau de neurones à convolutions

Farès Menasri^{1,2} – Nicole Vincent¹ – Emmanuel Augustin²

¹ Université PARIS DESCARTES - UFR de Mathématiques et Informatique
45 rue des Saints Pères, 75006 Paris - FRANCE

² A2iA SA
40 bis rue Fabert, 75007 Paris - FRANCE

{fm,ea}@a2ia.com , nicole.vincent@math-info.univ-paris5.fr

Résumé : *Une base de données de chiffres farsis (persans) manuscrits isolés de taille conséquente (60000 images en apprentissage et 20000 exemples en test) a été proposée [KHO 07].*

Dans cet article, nous proposons un système de reconnaissance de chiffres farsis isolés, et nous montrons que les techniques développés pour la reconnaissance de chiffres arabes sur la base MNIST s'appliquent également bien à la reconnaissance de chiffres indiens et farsis. En particulier, l'utilisation de réseaux de neurones à convolutions permet de s'affranchir de l'étape délicate du choix d'un jeu de primitives ad hoc ; le réseau de neurones apprend lui-même une sa propre extraction de primitives à partir des données. Nous vérifions également que l'utilisation de techniques de déformations élastiques pour accroître artificiellement la taille de la base d'apprentissage permet d'améliorer les performances du reconnaiseur.

Les performances rapportées (99,02% de reconnaissance) dépassent celles des systèmes déjà proposés sur la problématique de reconnaissance de chiffres farsis manuscrits.

Mots-clés : Reconnaissance de chiffres isolés, Chiffres farsis, Réseau de Neurones à convolutions.

1 Introduction

Les chiffres farsis sont utilisés principalement en Iran et en Afghanistan. Ils sont proches des chiffres indiens utilisés principalement dans les pays arabes du Moyen-Orient.

Comme le montre le tableau 1, les alphabets de chiffres arabes et farsis sont proches. Ils partagent les mêmes symboles, à l'exception du '4' et du '6'. Les symboles '0', '1', '2', '3', '7', '8', '9' sont rigoureusement identiques. Le 5 est très légèrement différent, mais compte tenu de la variabilité importante de l'écriture manuscrite, on peut s'attendre à ce que les '5' arabes et farsis soient quasiment identiques.

Compte tenu de la proximité avec l'alphabet de chiffres arabes, nous pouvons raisonnablement estimer qu'un système de reconnaissance appris sur une base similaire de chiffres arabes fournirait probablement des résultats très proches.

Dans cet article, nous proposons un système de reconnaissance des chiffres farsis. Ce système est entraîné sur une base de chiffres isolés proposée par H. Khosravi et E. Kabir

[KHO 07]. Cette base de données, qui est en quelque sorte l'Alter ego de la base MNIST¹, sera présentée dans la section 2. Le réseau de neurones à convolutions est présenté dans la section 3. Dans la section 4, nous décrivons une procédure de déformations élastiques utilisées pour accroître artificiellement la taille de la base d'apprentissage. Les résultats seront présentés dans la section 6. A titre de comparaison, nous donnerons également les performances d'autres reconnaiseurs de types perceptrons multi-couches, appris directement à partir de l'image initiale, ou après une extraction de primitives à l'aide de descripteurs de Fourier elliptiques.

Les conclusions seront données dans la section 7.

2 Base de données utilisée

Cette base de données a été rendue publique en février 2007 par H. Khosravi et E. Kabir [KHO 07]. Il s'agit d'une base conséquente de chiffres farsi isolés, d'une taille comparable à la base MNIST.

Elle est composée de 102352 chiffres extraits à partir d'environ 12000 formulaires, remplis par des étudiants et des personnels universitaires. Les documents sont scannés en 200dpi, binarisés, et la base est prédécoupée en un ensemble d'apprentissage (60000 exemples), un ensemble de test (20000 exemples), plus un troisième ensemble qui contient des données restantes (22352 exemples).

Les images sont stockées dans une base de données au format CDB. Du code C++ ou Matlab fourni par les auteurs permet d'extraire les chiffres isolés.

Le tableau 2 donne une liste de chiffres farsis manuscrits.

Les données sont fournies sous forme d'images bitonales de tailles variables. La taille maximale est de 51x62 pixels, tandis que la taille minimale est de 3x4 pixels. Une telle disparité s'explique par le fait que le caractère '0' est un point, et est donc beaucoup plus petit que les autres caractères.

Outre le cas particulier du zéro, les hauteurs des autres caractères de la base de sont pas normalisées, et leur taille peut varier du simple au double en fonction des exemples.

La manière dont sont stockées les données ne permet pas de retrouver les documents dont sont extraites les images. Il n'est donc pas possible de renormaliser leur taille en fonction des scripteurs. Ceci peut être gênant en particulier à cause de

¹<http://yann.lecun.com/exdb/mnist/index.html>

Europe	0	1	2	3	4	5	6	7	8	9
Monde Arabe (Moyen Orient)	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Iran/Pakistan (farsi)	٠	١	٢	٣	٤	٥	٦	٧	٨	٩

TAB. 1 – Chiffres utilisés en Europe, dans le monde Arabe, et en Iran/Pakistan.



TAB. 2 – Exemples de caractères farsi manuscrits.

la ressemblance des symboles '0' et '5', qui ne se différencient que par la taille : '0' est un point, tandis que '5' est un cercle. Dans certains cas, on aura une confusion réelle entre des images de ces deux classes. Ce problème est discuté dans la partie résultats (section 6).

Pour prendre en compte la variabilité de la taille de l'écriture, il conviendra donc soit de normaliser ces caractères en hauteur, soit d'extraire des primitives invariantes, soit d'entraîner un système robuste aux variations de hauteur de l'écriture. C'est cette troisième méthode qui sera mise en oeuvre. Les images sont simplement sous-échantillonnées de telle sorte que leur hauteur et leur largeur soient divisées par deux. Elles sont ensuite centrées dans des imageries de taille 32x32 en niveaux de gris.

Ces imageries de taille réduite et en niveaux de gris seront directement soumises en entrée d'un réseau de neurones à convolutions.

3 Réseau de neurones à convolutions

3.1 Remarques générales

Ce type de réseaux de neurones tire profit du partage des poids qui limite le nombre de coefficients à apprendre, et donc la complexité du modèle. Il peut donc être appliqué directement sur l'image, ce qui permet de s'affranchir de l'étape d'extraction de primitives. C'est un atout intéressant, compte tenu du fait que le choix du jeu de primitives est souvent l'un des facteurs limitants d'un système de reconnaissance : le réseau de neurones apprend automatiquement sa propre extraction de caractéristiques en fonction des données d'apprentissage.

Ce type de système est robuste aux translations, rotations, facteurs d'échelle, étirements / écrasements, et épaisseur des traits. Il propose également des capacités de robustesse exceptionnelles aux bruits : hachures, ratures, bruit poivre et sel [LEC].

Ce type de réseau de neurones en particulier donne d'excellents résultats sur la base de chiffres MNIST. La reconnaissance de chiffres isolés ou en contexte est une application pour laquelle ce type de réseaux est particulièrement bien adapté. Ils ont également été appliqués avec succès à des tâches de détection de visages [Osa 05, OSA 07, CHE 06], de plaques d'immatriculations [CHE 06], de la détection d'objets sous différents angles et différentes conditions d'éclairage [LEC 04], ou encore pour la commande automatique du

déplacement d'un robot d'exploration [LEC 06]. A chaque fois, le système de reconnaissance tire avantage de la capacité du réseau de neurones à apprendre sa propre extraction de caractéristiques en fonction des données d'apprentissage.

3.2 Système utilisé

Dans [LEC 98], LeCun et al. présentent une topologie de réseau de neurones à convolutions adapté à la reconnaissance de l'écriture manuscrite. Dans [SIM 03], P. Y. Simard et al. proposent une autre topologie de réseaux de neurones à convolutions.

En s'inspirant de ces travaux, nous proposons une autre topologie de réseaux de neurones à convolution, similaire à celle d'un LeNet5, mais pour lequel le dernier étage (classifieur à base de combinaison de gaussiennes) est remplacé par un perceptron multicouches à sorties softmax.

Le système prend en entrées des images en niveaux de gris de taille 32x32. Un tel réseau est entraîné par rétropropagation du gradient d'erreur. Il apprend sa propre extraction de caractéristiques conjointement à l'apprentissage de la tâche de classification.

Le réseau de neurones est appris de manière itérative par descente de gradient stochastique. Les exemples sont présentés en entrée du réseau de neurones et l'annotation correspond à la classe de sortie attendue (qui correspond à une valeur de sortie attendue de 1 pour cette classe, et 0 pour toutes les autres). Une fonction de transfert Softmax assure que les sorties du réseau de neurones sont homogènes à des probabilités. La fonction de coût utilisée est la divergence de Kullback-Leibler :

$$C_{KL} = \sum_i T_i \log\left(\frac{T_i}{S_i}\right)$$

où T_i correspond à la sortie désirée, et S_i la valeur de sortie réelle du neurone i .

4 Déformations élastiques

Souvent, dans les systèmes de reconnaissance, on travaille sur l'extraction de caractéristiques, de manière à rendre les primitives extraites robustes par rapport à la variabilité de l'écriture. De cette manière, on simplifie la tâche du reconnaissseur. Mais cette simplification se fait au prix d'une complexité accrue de la tâche d'extraction de primitives, qui risque d'être sur-adaptée pour un type de données particulier, et de mal se généraliser à d'autres données.

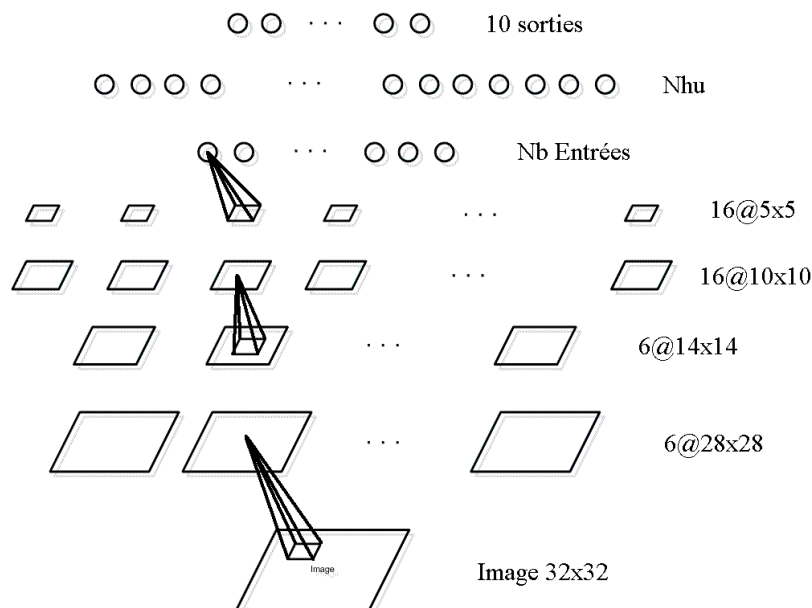


FIG. 1 – Système utilisé pour la reconnaissance des chiffres farsi. Topologie inspirée des travaux de [LEC 98] et [SIM 03]. L'image en niveaux de gris est soumise en entrée du système. Une alternance de couches de convolution et de sous-échantillonnage réalisent l'extraction de primitives, et un réseau de neurones de type Perceptron Multicouches à sorties softmax effectue la classification. Le système complet est entraîné par rétropropagation du gradient.

Le pendant de cette méthode consiste à entraîner le modèle sur davantage de données, de telle sorte que le reconnaissseur dispose d'une couverture d'exemples plus large, de manière à mieux prendre en compte la variabilité de l'écriture. C'est le sens des techniques qui consistent à augmenter artificiellement la taille de la base de données d'entraînement.

On pourrait craindre que le reconnaissseur soit sur-adapté pour prendre en compte ces types de variations pseudo-aléatoires. Mais en pratique, on constate que le fait de rajouter des données déformées visuellement réalistes permet effectivement d'améliorer la robustesse du reconnaissseur.

init.	1	2	3	4	5

TAB. 3 – Déformations élastiques. Cinq exemples d'images synthétiques obtenues pseudo-aléatoirement par déformations élastiques de l'image initiale.

Dans [SIM 03], Simard et al. décrivent un algorithme de déformation des données. Un champ de déformations est initialisé aléatoirement. Ce champ aléatoire est lissé par le biais d'une convolution à l'aide d'un filtre Gaussien. Ce champ de déformation est alors appliqué à l'image initiale pour obtenir une nouvelle image déformée. En fonction de la variance de la gaussienne et de l'intensité du champ de déformation aléatoire, on peut obtenir différents types de champs de déformations (voir figure 2).

Simard et al. proposent des paramètres de telle sorte que le modèle de distorsions élastiques offre un rendu proche de

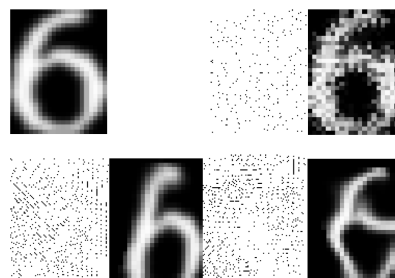


FIG. 2 – En haut à gauche : images initiale. En haut à droite : champ de déformation aléatoire non lissé, et image résultante. En bas : champs de déformations lissés et images résultantes. Figure extraite de [SIM 03]

la variabilité de l'écriture manuscrite. Cette méthode permet de générer des exemples de chiffres réalistes à partir des données déjà existantes (voir figure 3).

5 Descripteurs de Fourier elliptiques

A titre de comparaison, nous entraînons également un autre réseau de neurones de type perceptron multi-couches, dont les primitives sont extraites à partir des descripteurs de Fourier elliptiques.

Les descripteurs de Fourier elliptiques ont été introduits par F. P. Kuhl et C. R. Giardina [KUH 82] pour décrire les formes des contours fermés.

Une démonstration visuelle de leur fonctionnement est disponible sur [BRE].

Soit un contour de T pixel, dont chaque pixel a pour coordonnées $(x(t), y(t))$, avec $t \in \{0, 1, \dots, T - 1\}$.

$x : t \rightarrow x(t)$ et $y : t \rightarrow y(t)$ sont périodiques de période

T (le temps nécessaire pour tracer l'ensemble du contour à vitesse constante).

Ces signaux périodiques sont décomposables en séries de Fourier :

$$x(t) = A_0 + \sum_{n=1}^{\infty} \left[a_n \cos \frac{2n\pi t}{T} + b_n \sin \frac{2n\pi t}{T} \right]$$

$$y(t) = C_0 + \sum_{n=1}^{\infty} \left[c_n \cos \frac{2n\pi t}{T} + d_n \sin \frac{2n\pi t}{T} \right]$$

avec,

$$A_0 = \frac{1}{T} \int_0^T x(t) dt$$

$$C_0 = \frac{1}{T} \int_0^T y(t) dt$$

$$a_n = \frac{2}{T} \int_0^T x(t) \cos \frac{2n\pi t}{T} dt$$

$$b_n = \frac{2}{T} \int_0^T x(t) \sin \frac{2n\pi t}{T} dt$$

$$c_n = \frac{2}{T} \int_0^T y(t) \cos \frac{2n\pi t}{T} dt$$

$$d_n = \frac{2}{T} \int_0^T y(t) \sin \frac{2n\pi t}{T} dt$$

Différents niveaux d'approximation peuvent être obtenus en faisant varier le nombre d'harmoniques. L'approximation de rang N est obtenue comme :

$$x_N(t) = A_0 + \sum_{n=1}^N \left[a_n \cos \frac{2n\pi t}{T} + b_n \sin \frac{2n\pi t}{T} \right]$$

$$y_N(t) = C_0 + \sum_{n=1}^N \left[c_n \cos \frac{2n\pi t}{T} + d_n \sin \frac{2n\pi t}{T} \right]$$

Les coefficients a_n , b_n , c_n et d_n peuvent être utilisés comme primitives.

Ici, les paramètres correspondant aux 10 premiers rangs harmoniques sont conservés et sont utilisés comme primitives en entrées du réseau de neurones.

6 Expériences et résultats

La base de données utilisée est prédécoupée en un ensemble d'apprentissage (60000 images chiffres) et un ensemble de test (20000 images de chiffres). Les classes sont équiprobables.

Nous utiliserons un réseau de neurones à convolutions, qui est appliqué directement sur l'image et ne nécessite pas de phase d'extraction de caractéristiques. Nous appliquons également la technique proposée par Simard et al. pour augmenter la taille de l'ensemble d'apprentissage à l'aide de données pseudo-aléatoires obtenues en déformant les données d'entraînement existantes. Chaque image d'entraînement permet ainsi de générer 5 images supplémentaires.

Nous essayons différentes topologies de réseaux de neurones pour déterminer la meilleure configuration. Les perfor-

Entrées	couche cachée	base classique	base étendue
120	200	98.34	99.02
120	100	98.65	98.97
80	100	98.54	98.83
50	100	98.46	98.75

TAB. 4 – Résultats de reconnaissance : réseau à convolutions.

couche cachée	base classique	base étendue
50	95.78	96.41
100	96.55	97.05
150	95.82	97.04

TAB. 5 – Résultats de reconnaissance : perceptron multi-couches appliqué directement sur les images.

mances sont stables. Comme le montre le tableau 4, l'architecture dont le premier étage extrait un vecteur de 120 primitives et dont le perceptron multicouches contient 100 neurones sur la couche cachée permet d'obtenir les meilleures performances en utilisant la base d'entraînement classique. L'augmentation artificielle de la taille de la base d'entraînement permet d'améliorer sensiblement les performances. Le système avec 200 neurones sur la couche cachée devient légèrement meilleur avec le jeu de données étendues : les performances sont très proches, et se situent autour de 99% de reconnaissance sur la base de test.

A titre de comparaison, un réseau de neurones de type perceptron multi-couches a été entraîné sur les mêmes données, directement à partir de l'image (tableau 5), ou bien après une extraction de primitives à l'aide des descripteurs de Fourier elliptiques (tableau 6).

L'approche naïve qui consiste à appliquer un perceptron multi-couches directement sur l'image d'entrée donne les moins bonnes performances (cf. Tab. 5). L'extraction de primitives à l'aide des descripteurs de Fourier elliptiques se comporte mieux (cf. Tab. 6). Mais les réseaux à convolution sont l'approche qui permet d'obtenir les meilleures performances (cf. Tab. 4).

6.1 Comparaisons avec d'autres systèmes

A ce jour, à notre connaissance aucun taux de référence n'a été publié sur cette base de données pourtant à la fois publique et de taille conséquente. En revanche, à titre indicatif, nous pouvons indiquer les performances de systèmes évalués sur des bases privées de tailles plus petites. Ces taux sont donnés dans le tableau 8.

couche cachée	base classique	base étendue
200	97.58	97.56
300	97.64	97.78
400	97.61	97.68

TAB. 6 – Résultats de reconnaissance : perceptron multi-couches après extraction de primitives de type descripteurs de Fourier elliptiques (les 10 premiers rangs harmoniques sont utilisés).

	0	1	2	3	4	5	6	7	8	9
0	98.35	0.00	0.00	0.00	0.05	1.25	0.05	0.30	0.00	0.00
1	0.30	99.50	0.00	0.00	0.05	0.00	0.00	0.00	0.05	0.10
2	0.00	0.10	98.00	1.35	0.15	0.00	0.10	0.15	0.00	0.15
3	0.00	0.00	1.00	98.15	0.75	0.05	0.00	0.00	0.05	0.00
4	0.00	0.00	0.10	1.45	98.25	0.05	0.05	0.00	0.00	0.10
5	0.05	0.00	0.00	0.05	0.05	99.75	0.05	0.00	0.05	0.00
6	0.00	0.15	0.10	0.00	0.00	0.10	98.95	0.05	0.00	0.65
7	0.00	0.05	0.05	0.00	0.05	0.00	0.10	99.75	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.80	0.20
9	0.00	0.05	0.00	0.00	0.05	0.00	0.20	0.00	0.00	99.70

TAB. 7 – Matrice de confusion pour le cas 120 entrées et 200 neurones sur la couche cachées, entraîné sur la base de données étendue.

Système	Taux de reconnaissance
Notre système, sur la base [KHO 07]	99.02
[HOS 96]	92
[MOW 02]	91.81
[HAR 05]	97.6
[MOZ 05]	94.44
[ZIA 07]	98.41

TAB. 8 – Taux de reconnaissance de systèmes de reconnaissance de chiffres farsis.

6.2 Analyse d'erreurs

L'analyse de la matrice de confusion (table 7) permet de déterminer deux types de confusions :

- certains chiffres '0' sont reconnus comme des '5'
- des confusions entre les chiffres '2', '3' et '4'.

Des exemples de confusions sont donnés dans le tableau 9.

Les confusions entre '0' et '5' sont dues à l'absence de normalisation en hauteur entre les caractères. Sur un champ numérique complet, les '0' sont faciles à différencier des autres symboles. Il est possible de les reconnaître durant une phase de prétraitement du champ numérique, sur la base de règles très simples, puis d'utiliser un reconnaiseur plus élaboré pour différencier les autres symboles. Il n'est toutefois pas possible d'appliquer une telle approche sur cette base : elle ne contient que des images de chiffres isolés, et il est impossible de reconstituer les champs dont les images sont extraites.

La confusion entre les chiffres '2', '3' et '4' est plus problématique. Selon les scripteurs, on voit (tableau 9) que ces symboles sont parfois visuellement très proches. La décision est difficile, même pour un opérateur humain.

7 Conclusions

Ces expériences confirment l'intuition selon laquelle les réseaux à convolution, qui donnent de très bons résultats de reconnaissance sur la base de chiffres MNIST, pourraient avantageusement être appliqués à la reconnaissance de chiffres farsis sans aucune modification et sans nécessiter une phase d'extraction de primitives ad hoc.






Ces expériences confirment également le fait que les techniques de déformations élastiques pour augmenter artificiellement la taille de l'ensemble d'apprentissage s'appliquent

bien aux chiffres farsis.

Plusieurs systèmes ont été évalués par leurs auteurs sur des bases privées, ce qui empêche les comparaisons entre systèmes. La base de données proposée par H. Khosravi et E. Kabir [KHO 07] pourrait être utilisée à cet effet.

Références

- [BRE] BRECHBUHLER C., HO S., Reconstruction of a closed curve from its elliptic descriptor, <http://www.cs.unc.edu/styner/misc/FourierDemo/>.
- [CHE 06] CHEN Y.-N., HAN C.-C., WANG C.-T., JENG B.-S., , FAN K.-C., The Application of a Convolution Neural Network on Face and License Plate Detection, *ICPR '06 : Proceedings of the 18th International Conference on Pattern Recognition*, 2006, pp. 552–555.
- [HAR 05] HARIFI A., AGHAGOLZADEH A., A New Pattern for Handwritten Persian/Arabic Digit Recognition, *PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY*, vol. 3, 2005.
- [HOS 96] HOSSEINI H. M. M., BOUZERDOUM A., A combined method for Persian and Arabic handwritten digit recognition, NARASIMHAN V. L., JAIN L. C., Eds., *ANZIS*, IEEE, 1996, pp. 80-83.
- [KHO 07] KHOSRAVI H., KABIR E., Introducing a very large dataset of handwritten Farsi digits and a study on their varieties, *Pattern Recognition Letters*, vol. 28, n° 10, 2007, pp. 1133–1141, Elsevier Science Inc.
- [KUH 82] KUHL F. P., GIARDINA C. R., Elliptic Fourier feature of a closed contour, *Computer Vision, Graphics and Image Processing*, vol. 18, 1982, pp. 236-258.
- [LEC] LECUN Y., LeNet-5, convolutional neural networks, <http://yann.lecun.com/exdb/lenet/index.html>.

Classe réelle	Classe reconnue	Exemples
0	5	
2	3	
3	2	
3	4	
4	3	

TAB. 9 – Erreur de reconnaissance.

[LEC 98] LECUN Y., BOTTOU L., BENGIO Y., HAFNER P., Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, vol. 86, n° 11, 1998, pp. 2278–2324.

[LEC 04] LECUN Y., HUANG F. J., BOTTOU L., Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting, *cvpr*, vol. 02, 2004, pp. 97-104.

[LEC 06] LECUN Y., MULLER U., BEN J., COSATTO E., FLEPP B., Off-Road Obstacle Avoidance through End-to-End Learning, WEISS Y., SCHÖLKOPF B., PLATT J., Eds., *Advances in Neural Information Processing Systems 18*, pp. 739–746, MIT Press, Cambridge, MA, 2006.

[MOW 02] MOWLAEI A., FAEZ K., HAGHIGHAT A., Feature extraction with wavelet transform for recognition of isolated handwritten Farsi/Arabic characters and numerals, *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, vol. 2, 2002, pp. 923-926 vol.2.

[MOZ 05] MOZAFFARI S., FAEZ K., ZIARATBAN M., Structural Decomposition and Statistical Description of Farsi/Arabic Handwritten Numeric Characters, *ICDAR '05 : Proceedings of the Eighth International Conference on Document Analysis and Recognition*, 2005, pp. 237–241.

[Osa 05] OSADCHY M., MILLER M. L., LECUN Y., Synergistic Face Detection and Pose Estimation with Energy-Based Models, SAUL L. K., WEISS Y., BOTTOU L., Eds., *Advances in Neural Information Processing*

Systems 17, pp. 1017-1024, MIT Press, Cambridge, MA, 2005.

[OSA 07] OSADCHY M., CUN Y. L., MILLER M. L., Synergistic Face Detection and Pose Estimation with Energy-Based Models, *J. Mach. Learn. Res.*, vol. 8, 2007, pp. 1197–1215, MIT Press.

[SIM 03] SIMARD P., STEINKRAUS D., PLATT J. C., Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis, *ICDAR*, 2003, pp. 958-962.

[ZIA 07] ZIARATBAN M., FAEZ K., EZOJI M., Use of Legal Amount to Confirm or Correct the Courtesy Amount on Farsi Bank Checks, *ICDAR '07 : Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, 2007, pp. 1123–1127.