1. Tema de investigación

El reconocimiento de dígitos manuscritos es un tema que ocupa distintas áreas de conocimiento e investigación tanto en aprendizaje automático como en clasificación de patrones. Su importancia se debe al interés que genera la posibilidad de su aplicación práctica mediante la automatización de la comprensión del alto volumen de documentos impresos en empresas, gobiernos y una sociedad en expansión permanente. Para citar algunos ejemplos la gestión automática de correo postal [1] [2], procesamiento de cheques bancarios [3], ingreso manual de datos en formularios [4], registros gubernamentales y tarjetas de crédito impresas son aplicaciones de gran uso.

Los `Datos Abiertos`, `Open Data`, según el site Definicion Abierta [1] proyecto de la Fundación del Conocimiento Libre (Open Knowledge Foundation) [2], una organización sin fines de lucro, los define como:

- `Acceso y disponibilidad`: Los datos deben estar disponibles por completo, si existiese un costo asociado, este debe ser razonable, preferentemente accesible sin costo por Internet. Deben estar disponibles en un formato conveniente y modificable.
- `Reutilización y redistribución`: Los datos deben permitir su re-distribución y reutilización, incluyendo su comercialización ya sea individualmente o con entrecruzamiento y/o entremezclado con otras bases de datos.
- `Participación universal`: La licencia debe permitir que todos tengan la posibilidad de utilizar, reutilizar y redistribuir, sin ningún tipo de discriminación a personas o grupos.

Toda base de datos abierta debería estar acompañada de su correspondiente licencia, la misma puede exigir una mención a quienes generaron los datos ó reservarse el derecho de que cualquier trabajo derivado de la base de datos abierta mantenga la misma licencia, etc. Es similar al `software libre`, donde libre no significa `gratis` y que la libertad se extiende hasta donde decide el o los autores originales del trabajo. Estos datos, se presentan de fácil acceso, en gran cantidad de casos, y pese a que exista la posibilidad técnica de ser almacenados en una máquina para luego disponer de ellos, es la legalidad de su utilización la que no siempre está permitida.
En el caso particular la database de dígitos manuscritos MNIST, por ser utilizado en esta tesis, los derechos de autor le corresponden a Yann LeCun [3] y Corinna Cortes, bajo la licencia *Creative Commons Attribution-Share Alike 3.0*, son datos abiertos de libre uso con la condición de mencionar a sus autores en trabajos derivados.

[1] http://opendefinition.org/od/2.0/en/index.html
[2] https://okfn.org/
[3] http://yann.lecun.com/

En Argentina, la legislación ha comenzado a incluir el término **Dato Abierto** ó **Dato Público** como una manera de que los ciudadanos puedan disponer de ellos, e idealmente produzcan un valor agregado.
El Gobierno de la Ciudad de Buenos Aires ha dispuesto un sitio web con datos públicos [4] reglamentado por la Ley de Acceso a la Información Pública Nro. 104 [5] y la Ley de Protección de Datos Personales Nro. 1845, a la vez que ha definido Datos Abiertos como: *Significa poner información del Estado en un catálogo al alcance de todos, en formatos digitales, estándar y abiertos, siguiendo una estructura clara que permita su fácil comprensión y reutilización por parte de la ciudadanía.*. Otras ciudades de la Argentina como Bahia Blanca [6] y Datos Abiertos Misiones [7], tienen su portal de Datos Públicos, lo que constituye una herramienta efectiva.

En la Ciudad de Buenos Aires, existe un tipo de dato público, que está disponible y que contiene los resultados electorales de la Ciudad de Buenos Aires, Argentina. Al finalizar cada acto eleccionario los telegramas son completados manualmente por las autoridades de cada mesa durante las elecciones de diputados y senadores en Argentina, Buenos Aires, Capital Federal. Estos documentos son transcriptos por data entries y posteriormente escaneados y publicados ambos, los escaneados y los tipeados, en web con acceso libre.

Los telegramas manuscritos disponibles por la Dirección Nacional Electoral [8], son resultados electorales y en virtud de la disposición 408/2013, lo dispuesto en el Anexo I del Decreto Nr. 682 del 14 de mayo de 2010, se aprueba la Directiva De Datos Públicos Abiertos Para La Administracion Electoral [9]. Esta disposición establece que los resultados electorales deberán adecuarse progresivamente a la directiva de permitir la interacción con los ciudadanos.

El presente trabajo de tesis de maestría aborda el reconocimiento automático de dígitos manuscritos y su clasificación con datos electorales del 27 de Octubre de 2013. Finalmente, el trabajo compara el tipeado manual realizado por data entries de la empresa contratada para dicha tarea, contra los números reconocidos en forma artificial.

En resumen, esta tesis propone el reconocimiento automático de dígitos manuscritos utilizando técnicas de análisis de imágenes y minería de datos a partir de imágenes de planillas manuscritas provinientes de datos públicos.

[4] http://data.buenosaires.gob.ar/about
[5] http://www.cedom.gov.ar/es/legislacion/normas/leyes/ley104.html
[6] http://bahiablanca.opendata.junar.com/home/
[7] http://www.datos.misiones.gov.ar/
[8] http://www.resultados.gob.ar/inicio.htm
[9] http://www.infoleg.gob.ar/infolegInternet/anexos/220000-224999/221756/norma.htm

Referencias

[1] Le Cun, B. B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems.

[2] Tay, Y. H., Lallican, P. M., Khalid, M., Viard-Gaudin, C., & Kneer, S. (2001). An offline cursive handwritten word recognition system. In TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology (Vol. 2, pp. 519-524). IEEE.

[3] Palacios, R., & Gupta, A. (2008). A system for processing handwritten bank checks automatically. Image and Vision Computing, 26(10), 1297-1313.

[4] Van der Zwaag, B. J. (2001). Handwritten digit recognition: A neural network demo. In Computational Intelligence. Theory and Applications (pp. 762-771). Springer Berlin Heidelberg.

[5] Gustav Tauschek (1935). Reading machine. US Patent 2.026.330.

[6] P. W. Handel (1933), Statistical machine. US. Patent 1.915.993.

[7] Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. Proceedings of the IEEE, 80(7), 1029-1058.

[8] M. Costa, E. Filippi and E. Pasero (1994), "A Modular cyclic Neural Network for character recognition", Proceedings of the INNS World Congress on Neural Networks (WCNN '94), S. Diego (CA), Vol 3, June 5-9, pp 204-210

[9] Guyon, I., Haralick, R. M., Hull, J. J., & Phillips, I. T. Data sets for OCR and document image understanding research. Handbook of character recognition and document image analysis, 779-799. 1997

[10] LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H. & Vapnik, V. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. Neural networks: the statistical mechanics perspective, 261, 276.

[11] LeCun, Y., Cortes, C., & Burges, C. J. (1998). The MNIST database of handwritten digits.

[12] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.