ICDAR2017 Competition on Information Extraction in Historical Handwritten Records

Alicia Fornés*, Verónica Romero†, Arnau Baró*, Juan Ignacio Toledo*,
Joan Andreu Sánchez†, Enrique Vidal†, Josep Lladós*
**Computer Vision Center - Computer Science Department, Universitat Autònoma de Barcelona, Spain*
Email: {afornes,abaro,jitoledo,josep}@cvc.uab.es
†PRHLT Research Center, Universitat Politècnica de Valencia, Spain
Email: {vromero,jandreu,evidal}@prhlt.upv.es

Abstract—**The extraction of relevant information from historical handwritten document collections is one of the key steps in order to make these manuscripts available for access and searches. In this competition, the goal is to detect the named entities and assign each of them a semantic category, and therefore, to simulate the filling in of a knowledge database. This paper describes the dataset, the tasks, the evaluation metrics, the participants methods and the results.**

1. Introduction

The extraction of relevant information from historical handwritten document collections is one of the key steps in order to make these manuscripts available for access and searches. In this context, instead of handwriting recognition [3], understood as pure transcription, the objective is to move towards document understanding. Concretely, the aim is to detect the named entities and assign each of them a semantic category, such as family names, places, occupations, etc. A typical application scenario of named entity recognition is demographic documents, since they contain people's names, birthplaces, occupations, etc. In this scenario, the extraction of the key contents and its storage in databases allows the access to their contents and envision innovative services based in genealogical, social or demographic searches. Lately, the interest of the document image analysis community in document understanding, named entity recognition and semantic categorization is awaking, and several techniques based on Hidden Markov Models (HMMs) [5], Bidirectional Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [1] and Convolutional Neural Networks (CNNs) [9] have been proposed.

With this competition [1], we aim to foster the research in this field and offer a benchmark for the research community. This competition will remain open and continuous, so that researchers can upload their new results at any time.

The rest of this paper is organised as follows. First, we describe the dataset in Section 2, the tasks in Section 3, and the participants' methods in Section 4. Then, we describe the evaluation metrics and discuss the results in Section 5.

2. The Esposalles database

For this competition we have used 125 pages of the Esposalles database [2], [4]. This database consists of historical handwritten marriages records from the Archives of the Cathedral of Barcelona. The pages we used correspond to the volume 69, written in old Catalan by one single writer in the 17th century. Each marriage record (see Figure 1) contains information about the husbands occupation, place of origin, husbands and wifes former marital status, parents occupation, place of residence, geographical origin, etc.

The structure of the marriage record tends to follow a regular expression. Some anchor words (in bold) separate the different persons, as follows:

$< husband >$ fill de (son of) $< husband'sfather >$ y (and) $< husband'smother >$ ab (with) $< wife >$ filla de (daughter of) $< wife'sfather >$ y (and) $< wife'smother >$.

In some cases, other persons may appear in the record. For example, when a widow is married again, the record may include information on the former husband. In those cases, the information of the wife's parents usually disappears:

$< husband >$ fill de (son of) $< husband'sfather >$ y (and) $< husband'smother >$ ab (with) $< wife >$ viuda (widow) $< wife'sformerhusband >$.

It must be noted that the above structures are usually followed, but in some cases, they present variations.

3. Tasks

The objective is to extract information from the records. Concretelly, the task is to recognize the named entities, such as names, surnames, places, occupations, etc. In order to foster the participation in this competition, we have simplified the number of semantic classes existing in the database. For example, the place of residence, geographical origin, etc. have been simplified to the same semantic class *location*.

For this competition, we have manually labelled the marriage records with semantic information at word level. The lines and the records in this dataset have been also manually annotated. In this way, each line is associated to its corresponding record.

[1] http://www.cvc.uab.es/5cofm/competition/

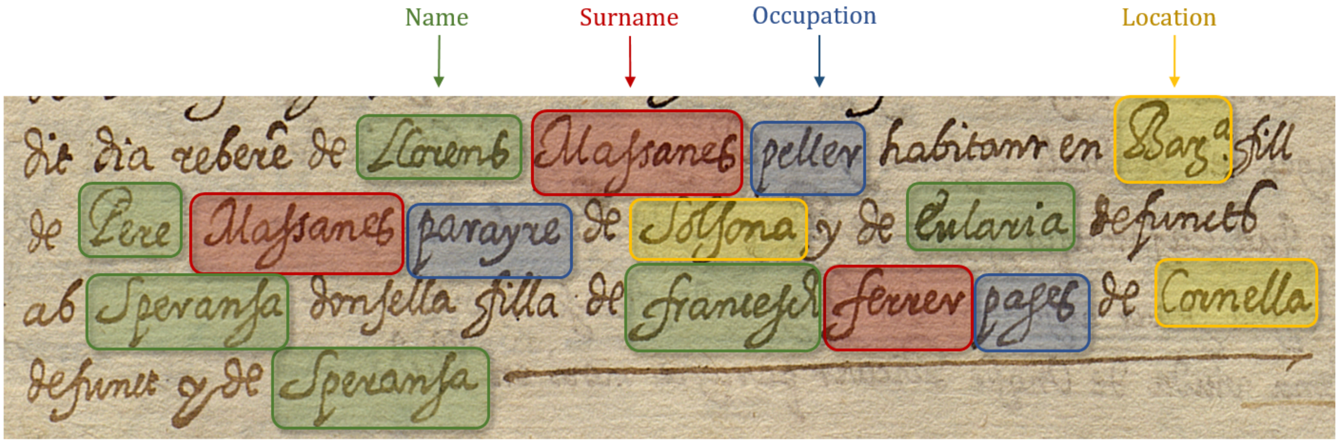


Figure 1. Example of a marriage record.

The training and test sets are composed of:

- Training set: 100 pages, 968 marriage records.
- Test set: 25 pages, 253 marriage records.

For each marriage record, we provide:

- Images of segmented text lines.
- Images of segmented words.
- Text files with the corresponding transcription.
- Text files with the corresponding categories: name, surname, occupation, location, civil state.
- Text files with the corresponding person: husband, husband's father, husband's mother, wife, wife's father, wife's mother, other-person (a different person from the ones mentioned before, for example, a former husband).
- A CSV file with the list of transcriptions, categories and associated persons.

The dataset is grouped into records. There is one folder per record, composed of the folders *lines* and *words*, and the corresponding CSV file. The CSV file only contains the relevant words, i.e. the named entities. This means that only words with an associated category (e.g. names, locations, etc.) will appear in the CSV file. Note that the final goal is to simulate the filling in of a knowledge database.

All TXT files are provided in correspondence, that is, each word in the marriage record will have associated its category (just one category per word). This information has been manually checked for avoiding inconsistencies, but take into account that some names and locations are composed of several words. For those non-relevant words (e.g. conjunctions, prepositions, verbs, etc.) the category will be *other* and the person will be *none*. An example of the provided ground-truth is shown in Figure 2.

Participants must provide, for each record, the CSV file with the transcription of the relevant words (i.e. named entities) and their semantic category. However, providing the person associated to each category is optional. Therefore, participants can decide in which track they would like to participate, either:

- **Track 1 - Basic**. The CSV must contain the transcription and the semantic category (name, surname, occupation, etc.).
- **Track 2 - Complete**. The CSV must contain the transcription, the semantic category and the person (husband, wife, wife's father, etc.).

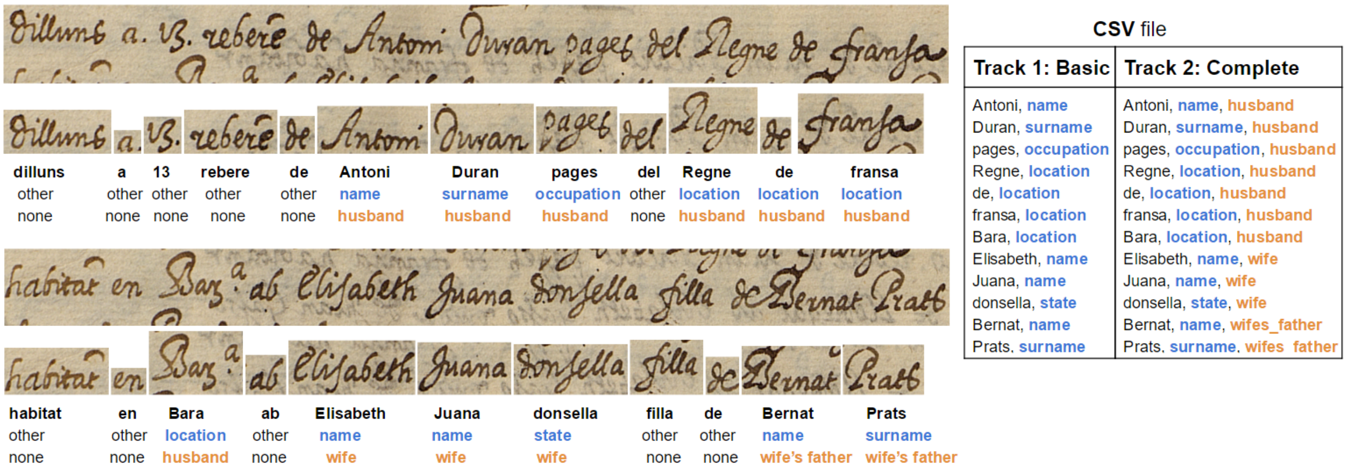
4. Methods

This section is devoted to describe the participants' methods and the two baseline methods. Three different research teams have participated, submitting a total of 5 methods. The first two teams are from Shenzen University, whereas the third team is from Rostock University.

4.1. Team 1: Hitsz-ICRC-1

- **Participants' team**: Xiangping Wu, Qingcai Chen, Linlin Wang, Qing Zhang.
- **Organization**: Harbin Institute of Technology Shenzhen Graduate School, Intelligent Computing Research Center. China.
- **Track**: Complete (transcription, category, person).
- **Segmentation level**: Word.

Method: CNN based Bi-gram method for segment-free liaison handwriting recognition and NER tagging. This method is divided into two parts: character recognition and named entity recognition. In the handwritten old Catalan text recognition stage, we present a novel, segmentation-free, word-wise character recognition method without any external linguistic knowledge. In this method, the position information of each character is converted into a vector. A kind of bi-gram model is then constructed and integrated into the convolution neural network for training. The whole process of character recognition consists three steps: (1) data pre-processing; (2) model training; and (3) model running. In the first step, we normalize the color word image to the size 100x200 and add a terminator '*' at



Track 1: Basic	Track 2: Complete
Antoni, name	Antoni, name, husband
Duran, surname	Duran, surname, husband
pages, occupation	pages, occupation, husband
Regne, location	Regne, location, husband
de, location	de, location, husband
fransa, location	fransa, location, husband
Bara, location	Bara, location, husband
Elisabeth, name	Elisabeth, name, wife
Juana, name	Juana, name, wife
donsella, state	donsella, state, wife
Bernat, name	Bernat, name, wifes_father
Prats, surname	Prats, surname, wifes_father

Figure 2. Example of the Ground-truth provided for a marriage record.

the end of each word. We compute the statistics of the bi-gram combination of all the characters on the training set. Since the numbers are not combined with the letters, we selected 2560 bi-gram combinations from 60 characters (59 primitives and 1 terminator) as the training classes. Next, we take into consideration the spatial location of the bi-gram inside the word, select 14 positions and convert each position to a multi-dimensional random vector. The random vector of the position information is only generated once.

In the second step, we use convolution neural network (CNN) and combine location information to build a system which, given an image, produces a prediction of the image transcription without constructing any attribute features. The network is trained using the aggregated sigmoid cross-entropy (logistic) loss and a learning rate of 0.01. In the final step, given an image and a location vector, it is run through the network. Then the network outputs the prediction results corresponding to the location of the image transcription. In the running step, we output the recognition results for 14 positions of each word, and then remove the first occurrence of the terminator and the following characters. When the predictions are in conflict, corrections are applied for post processing, according to the probability of the bi-gram frequency statistic only for the training set. For example, if the probability of 'er' is greater than the probability of 'eu', we choose 'er' as the final prediction result.

In the named entity identification stage, we simply use the CRF sequence tagging method via the CRF++ tool box. We first predict the category based on the record and the first template. And then predict the person based on another template and the record of the transcript and the category predicted in the previous step.

4.2. Team 2: Hitsz-ICRC-2

- **Participants' team**: Xiangping Wu, Qingcai Chen, Jinghan You.
- **Organization**: Harbin Institute of Technology Shenzhen Graduate School, Intelligent Computing Research Center. China.
- **Track**: Complete (transcription, category, person).
- **Segmentation level**: Word.

Method: Resnet based uni-gram method for segment-free liaison handwriting recognition and NER tagging.
This method is similar to the previous one, but instead of CNNs, a Resnet is used. The method is divided into the character recognition and named entity recognition. We present a novel, segmentation-free, word-wise character recognition method without any external linguistic knowledge. Here, the position information of each character is converted into a vector. A kind of uni-gram model is then constructed and integrated into the residual neural netwok for training. The character recognition process consists of: (1) data pre-processing; (2) model training; and (3) model running. In the first step, we normalize the color word image to the size 100x200 and add a terminator '*' at the end of each word. Second, model building and training. The first part of the recognition model draws on the resnet network to extract the feature from the input image and then a feature vector is generated. At the same time, we randomly generate a multi-dimensional vector for each location. Next, we combine the eigenvectors generated by the resnet network and the randomly generated multi-dimensional position vectors into a new feature vector. At the end of the network we added a fully connected neural network with a hidden layer and the dropout of 0.5 is used. The total number of network output layer units is 60, including 59 classes of Catalan basic characters and a terminator '*'. Third, model prediction. According to the statistics for the training set, we calculated the length of the longest word with the terminator. Then, the predicted length of the word in the test set is set to 15 to ensure that the end of the long word can be identified. Finally, we remove all the terminator to get the word predictions. This character recognition method does not depend on external language information such as dictionaries. The main contributions of the location information is to guide the resnet network to automatically learn the knowledge

of segmenting characters and to identify the corresponding location of the characters.

The named entity identification stage is the same as in the previous method. We simply use the CRF sequence tagging method via the CRF++ tool box. We first predict the category based on the record and the first template. The person is predicted based on another template, the record of the transcript and the category predicted before.

4.3. Team 3: CITlab ARGUS

- **Participants' team**: Tobias Strauß, Max Weidemann, Johannes Michael, Gundram Leifert, Tobias Grüning, Roger Labahn.
- **Organization**: University of Rostock (Institute of Mathematics). Germany.
- **Track**: Complete (transcription, category, person).
- **Segmentation level**: Line.

Method 1: CITlab-ARGUS-1 (without OOV).

The training data is divided into a training set (2790 line images) and a validation set (280 line images). Several normalization methods such as contrast, size, slant and skew normalization are applied. These preprocessed line images serve as input for the optical model, a recurrent neural network (RNN) (layer from input to output: conv, conv, lstm (256 cells), conv, lstm (512 cells)) trained by CTC (150 epochs of 5000 noisy line images each). To enlarge input variety in the line images, we use data augmentation on line images. The output of the optical model are probabilities for each character at each position in the image collected in a matrix. The various output matrices for one record (which represent the lines) are glued together to one single matrix. We define regular expressions to extract the required information from this matrix. This is done in two steps: First, we segment the matrix into regions of interest: regions containing information about the husband, the husbands parents, the wife or the wife's parents. These regions are matched against a valid combination of dictionary items in a second step (this means that Out Of Vocabulary words cannot be recognized).

Method 2: CITlab-ARGUS-2 (with OOV).

This method is the same as the one explained before, a RNN-LSTM with CTC. In this case, the system is able to recognize Out of Vocabulary (OOV) words. For the name fields additional OOV words are allowed if the dictionary items do not fit.

Method 3: CITlab-ARGUS-3 (with OOV).

This method is the same as the previous one (CITlab-ARGUS-2), but with a different network setting. Here, the optical model is a recurrent neural network (layer from input to output: conv, conv, blstm (512), conv, blstm (512 cells), blstm (512 cells)) trained by CTC (150 epochs of 5000 noisy line images each). As in the method described above, for the name fields, additional OOV words are allowed if the dictionary items do not fit.

4.4. Baseline 1 - CNNs

- **Track**: Complete (transcription, category, person).
- **Segmentation level**: Word.

This baseline method is based on Convolutional Neural Networks (CNNs). We divide the 100 pages of available training data into 90 pages (28346 word images) for train and 10 pages (3155 word images) for validation. This data is used to train two different neural network models. The first model is trained to perform the semantic categorization. The network is a relatively small CNN like the one described in [9] that can accept word images and outputs the semantic category of each word. For this competition, we used the same parameters as the ones mentioned in [9].

The second model is used to perform the transcription. In this case the model has two very diferentiated parts; the first part is a CNN like the one in [7] that embeds small windows of text into the PHOC space. The second part is a two-layer BLSTM network that performs the sequence recognition and outputs the transcription. This method is described in detail in [8]. Both methods were trained using 'early stopping', that is, to keep training until no improvement in validation accuracy is observed for a certain number (20) of epochs.

Finally, a parser is used to assign the person to the categories. We make use of the anchor words to distinguish the persons. For example, the keyword 'ab' marks the starting of the information concerning the 'wife'. The keyword 'fill' separates the husband from his parents, while 'filla' separates the wife from her parents. And the word 'y' is used to separate the father from the mother.

4.5. Baseline 2 - HMMs

- **Track**: Complete (transcription, category, person).
- **Segmentation level**: Line.

This baseline system is based on Hidden Markov Models (HMMs) and a category based n-gram model for language modeling. Then a Grammatical Inference technique known as MGGI has been used to improve the semantic accuracy of the category-based language model as described in [6]. In MGGI, a-priory knowledge is used to label the words of the training strings in such a way that a simple bigram can be trained from the transformed strings. The knowledge used allows the MGGI to produce a language model which captures important dependencies of the language underlying in the handwritten records considered.

The line images were preprocessed and a sequence of feature vectors based on the gray level of the image was obtained for each image. Since we carried out experiments at license level, the lines of the test set were concatenated into licenses. The characters were modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. These models were estimated using the Baum-Welch algorithm. For decoding we used the Viterbi algorithm.

5. Evaluation Results

5.1. Metrics

The evaluation is done at marriage record level. Since the focus of the competition is on information extraction, the semantic label is prioritized. This means that if the semantic label of a word is incorrect, the transcription is not taken into account. Contrary, if the semantic label is correct, then the Character Error Rate (CER) is used to evaluate the transcript. Concretelly, for each semantically labeled word:

- Track Basic. If the $category$ is incorrect, then the score is 0. Otherwise, compute the CER on the transcription.
- Track Complete. If the $category$ and $person$ are incorrect, then the score is 0. Otherwise, compute the CER on the transcription.

Note that the score at category level in both tracks is the same, so these values are directly comparable.

The procedure is the following. First, we check that the submissions are syntactically correct, that is, that one CSV file is provided for each record, that it has the right number of comma separated values and that all the categories, person and record id's are valid. We define the concept of semantic label as $category$ in the basic track and the concatenation of the $category$ and $person$ in the complete track.

For each semantically labeled word in each record, we retrieve two list of transcriptions: one from the submission and another one from the groundtruth. The CER is calculated for each pair of submission and ground-truth words. We calculate the CER as the Levenshtein Distance between the two words, normalized by the length of the longest transcription in order to have a value between 0 and 1.

Afterwards, the best alignment is determined with Dynamic Time Warping, and we calculate the average CER for that labeling. The accuracy score for each labeling is calculated as 1-CER, then we calculate the record accuracy as the average of the labeling accuracies found in the record. The final score is the average of all the record accuracies.

In addition to this final score, the average scores for each one of the categories are also computed. For better visualization, all these values are normalized between 0-100.

5.2. Results

Table 1 shows the average score of the basic and complete tracks. Although all teams have participated in the track complete, for the sake of completeness, we also show their results on the track basic. In the table, the first rows show the results of the methods that have used segmented words, whereas the rows in the bottom show the methods that used text lines. It must be noted that the difficulties when recognizing words or lines differ. For this reason, although the metrics are the same, their results are not directly comparable. From the above results, one can observe that the bests methods are the Hitsz-ICRC-2 for segmented words, and the CITlab-ARGUS-2 for the segmented lines.

TABLE 1. Average Score. Tracks Basic and Complete. Values between 0-100%.

Method	Segmentation	Track Basic Average Score	Track Complete Average Score
Baseline-CNN	Word	79.40	70.18
Hitsz-ICRC-1	Word	87.56	85.72
Hitsz-ICRC-2	Word	94.16	**91.97**
Baseline-HMM	Line	80.24	63.08
CITlab-ARGUS-1	Line	89.53	89.16
CITlab-ARGUS-2	Line	91.93	**91.56**
CITlab-ARGUS-3	Line	91.61	91.17

TABLE 2. Results for Categories. Track Basic. Values between 0-100%.

Method	location	occupation	state	name	surname
Baseline-CNN	66.23	86.25	97.68	83.01	65.25
Hitsz-ICRC-1	89.33	91.03	97.82	91.82	69.13
Hitsz-ICRC-2	94.90	93.76	95.35	95.67	91.19
Baseline-HMM	78.73	90.22	93.79	81.06	60.13
CITlab-ARGUS-1	87.61	92.64	97.43	94.36	76.53
CITlab-ARGUS-2	88.40	93.07	97.54	95.13	85.76
CITlab-ARGUS-3	87.30	92.95	97.19	95.08	85.81

Tables 2 and 3 show the average scores computed at category level. In this case, instead of computing the average of all records in the database, we compute the average for each category. Note that some combinations of person-category are very rare in the database (e.g. the surname of the husband's mother, the civil state of other persons), and consequently, many methods fail in their recognition, probably due to the insuficient examples in the training set. Contrary, the surname of the wife's mother never appears in the database (for this reason, in the table it is represented by the symbol -), however, some methods erroneously detect such a semantic label, and therefore, their score in this category is 0.

From these tables, it can be observed that the categories with fewer number of different words (i.e. vocabulary size) tend to have a higher performance. For example, the civil state and the occupation have a small and limited vocabulary (the amount of different civil states and occupations is small), and therefore, the named entity detector is more accurate. Contrary, categories with large vocabulary, or even with many out of vocabulary words (such as surnames), tend to obtain a lower performance.

6. Conclusions

In this competition on information extraction in historical documents, we have aimed to raise the interest in semantic recognition and categorization, as a first step towards the understanding of handwritten documents. We strongly believe that it is an interesting problem for the community, 5.

TABLE 3. RESULTS FOR CATEGORIES. TRACK COMPLETE. VALUES BETWEEN 0-100%.

Method	wifes-mother surname	wife name	husband occupation	wife occupation	husband surname	wife surname	husband state	husbands father location	husbands father name	husbands father occupation	husbands father surname	wife state	husbands-mother name
Baseline-CNN	0	88.50	87.57	0.27	65.38	6.02	96.92	20.16	75.31	88.11	60.93	97.00	92.19
Hitsz-ICRC-1	-	92.80	90.62	94.47	69.14	1.59	96.92	15.00	91.56	91.50	65.98	97.50	92.94
Hitsz-ICRC-2	0	94.99	93.29	93.65	91.11	0	96.92	16.00	94.31	92.35	85.93	95.70	96.72
Baseline-HMM	0	87.14	75.32	38.37	47.58	8.33	95.08	15.43	64.03	69.45	41.37	91.72	77.82
CITlab-ARGUS-1	-	97.66	90.01	90.68	80.68	8.81	92.54	63.85	92.97	89.74	73.64	97.13	95.94
CITlab-ARGUS-2	-	98.49	88.49	91.43	88.85	36.57	92.42	78.61	94.28	92.07	86.57	97.13	96.17
CITlab-ARGUS-3	-	98.38	88.12	91.43	89.14	41.67	93.94	74.72	92.57	91.83	85.05	96.74	95.82

Method	wifes father location	husbands mother surname	wifes father name	other person name	wifes father occupation	other person surname	wifes father surname	other person state	husband location	wifes mother name	wife location	husband name
Baseline-CNN	60.52	0	70.99	69.61	84.34	61.58	61.81	0	71.49	88.17	2.73	79.78
Hitsz-ICRC-1	75.18	0	88.17	93.55	88.90	69.66	67.43	0	88.85	92.99	86.52	92.83
Hitsz-ICRC-2	82.56	0	91.53	96.54	90.94	91.13	88.22	0	94.86	97.66	90.38	94.69
Baseline-HMM	47.14	0	58.92	42.56	56.54	29.89	38.48	0	74.29	62.35	31.50	71.19
CITlab-ARGUS-1	89.10	0	93.07	94.09	90.57	75.33	77.38	0	89.76	96.35	67.69	95.01
CITlab-ARGUS-2	89.29	0	94.42	93.93	89.17	88.06	87.43	0	90.42	95.90	66.73	96.10
CITlab-ARGUS-3	87.18	0	93.48	94.85	88.47	88.35	87.80	0	89.64	94.61	66.31	96.22

because instead of pure transcription, we need to investigate more intelligent reading systems, able to extract the information contained in a document and fill in a database. This competition aims to serve as a benchmark for the research community. Moreover, we plan to keep the competition open and continuous. This means that researchers can upload new results at any time, since the platform will compute the metrics for these new methods and show the results.

Acknowledgments

This work has been partially supported by the Spanish projects TIN2015-70924-C2-1-R and TIN2015-70924-C2-2-R, the European Union's H2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943), the Ramon y Cajal Fellowship RYC-2014-16831, the grant 2013-DI-067 from the Secretaria dUniversitats i Recerca del Departament dEconomia i Coneixement de la Generalitat de Catalunya, and the CERCA Programme/Generalitat de Catalunya.

References

[1] C. Adak, B. B. Chaudhuri, and M. Blumenstein, "Named entity recognition from unstructured handwritten document images," in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*. IEEE, 2016, pp. 375–380.

[2] D. Fernández-Mota, J. Almazán, N. Cirera, A. Fornés, and J. Lladós, "Bh2m: The barcelona historical, handwritten marriages database," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 256–261.

[3] V. Frinken and H. Bunke, "Continuous handwritten script recognition," in *Handbook of Document Image Processing and Recognition*. Springer, 2014, pp. 391–425.

[4] V. Romero, A. Fornés, N. Serrano, J. Sánchez, A. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The esposalles database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognition*, vol. 46, no. 6, pp. 1658–1669, 2013.

[5] V. Romero, A. Fornés, E. Vidal, and J. A. Sánchez, "Using the mggi methodology for category-based language modeling in handwritten marriage licenses books," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 331–336.

[6] ——, "Information extraction in handwritten marriage licenses books using the mggi methodology," in *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings*, 2017, pp. 287–294. [Online]. Available: https://doi.org/10.1007/978-3-319-58838-4_32

[7] S. Sudholt and G. A. Fink, "Phocnet: A deep convolutional neural network for word spotting in handwritten documents," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 277–282.

[8] J. I. Toledo, S. Dey, A. Fornés, and J. Lladós, "Handwriting recognition by attribute embedding and recurrent neural networks," in *Proc. of the ICDAR 2017*, 2017.

[9] J. I. Toledo, S. Sudholt, A. Fornés, J. Cucurull, G. A. Fink, and J. Lladós, "Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2016, pp. 543–552.