Language Identification from Handwritten Documents

Luc Mioulet[†], Utpal Garain[*], Clément Chatelain[‡], Philippine Barlas[†] and Thierry Paquet[†]
[†]Laboratoire LITIS - EA 4108, Universite de Rouen, FRANCE 76800
[*] CVPR Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India.
[‡]Laboratoire LITIS - EA 4108, INSA Rouen, FRANCE 76800

Abstract—**This paper presents a novel approach for language identification in handwritten documents. The approach is based on script identification followed by character recognition. BLSTM-CTC based handwriting recognizers are used and the OCR output is fed to a statistical language identifier for detecting the language of the input handwritten document. Documents in two scripts (Latin and Bengali) and four languages (English, French, Bengali and Assamese) are considered for evaluation. Several alternative frameworks have been explored, effects of handwriting recognition and text length on language detection have been studied. It is observed that with some empirical restrictions it is very much possible to achieve more that 80% language detection accuracy and based on the current research practical systems can be designed.**

Keywords*—*Script, Language, Handwritten documents, Character recognition, BLSTM-CTC, Language identification.

I. Introduction

Script and language are two different properties of a text document. Script refers to the alphabet (or character set) used to write the documents whereas language of the document refers to the usage of the characters to form valid words and sentences that obey a particular lexicon and grammar. A script may be used for writing many language, for cultural and historical reasons each script refers to a set of languages which are usually written using that script. Conversely, a language often refers to a conventional script in which it is normally written. For example, languages like English, French, Dutch, German, etc. are written using the Latin script, whereas Hindi, Sanskrit, Marathi, etc. are written using Devanagari script.

So far, the document image analysis (DIA) researchers have done considerable amount of research on automatic script identification [1]. After successful development of OCR systems for many different scripts, script identification became a necessity. The main task in this problem is to identify the script in which a document is printed. Once the script of the document is identified, respective OCR engine can be used to identify characters. Attempts have also been made to identify different scripts in multi-script documents [2]. Though, most research works on script identification are restricted to printed documents only, and a few of them have been extended to handwritten documents too [3].

Though script identification has been widely explored in DIA, researchers have paid little attention to identify the language of a document. Office automation being the main goal, language identification was not needed earlier as the main interest was to get editable versions of text images and for this purpose, script identification followed by character recognition was sufficient. However, with the newer needs, language identification of a document is required for many applications like machine translation (MT), information extraction (IE), etc.

Among the previous works on language identification the research by Spitz [4] is worth mentioning. Printed documents are considered, two scripts (Latin-based and Han-based) and for each script several languages are considered. At first scripts are recognized and then within that script a specific language is identified. Character shape codes were used as features for such identification. Another notable work is due to Hochberg et al. [6] where language has been identified from handwritten documents. Six scripts and eight languages are considered and several connected component based features are used for script and language identification. Later on, several other works refer to language identification but basically these works addresses script recognition only as different languages under a script are not considered in these works (e.g., [3], [5]).

What essentially comes out is that the script and language though refer to two completely different attributes of a document have been recognized mostly by the same shape feature based approach. In this paper, we advocates that shape feature based approach is definitely suitable for script recognition because script refers to character shapes but such an approach may not be suitable for language identification as many languages may refer to the same set of characters. However, previously it was difficult to use any character recognition based approach for language identification from handwritten documents as the accuracy of handwriting recognition was indeed very poor, when no lexicon or language model can help the recognition process.

This paper investigates a novel framework for language identification from handwritten documents. Let us consider a practical situation where a handwritten note in some script (assume that the script is Latin and the language is French) is available and we want to get it translated in English, the language we understand. For realizing this application we propose the following four-step framework: (i) script identifier to identify that the script is Latin, (ii) Latin handwritten OCR (HOCR) engine, (iii) language identification to identify that the language is French, (iv) MT system for French to English translation. DIA community can provide tools for the first two steps and natural language processing (NLP) community can provide the latter two tools (i.e., language identifier and MT) to realize the overall application.

This work investigates the handshaking of the above two communities which was not explored before. Apparently it seems that the handshaking is straight forward: the output from the OCR engine will go as the input to the language identifier for identification of the language. However, there are several issues that need attention. The language identifiers [7] available with NLP community are developed to identify language from clean text. We do not know how these tools behave for OCR'd data. Secondly, once a script is identified, behaviour of the corresponding HOCR will not be the same for all the languages written using that script. Today's successful HOCRs are heavily dependent on the underlying language model and as a result, there is as such no HOCR which is known as Latin HOCR. Rather we talk about French HOCR, English HOCR, etc. because they use corresponding language models and lexicons. Therefore, even though we know that the note is written in Latin script, we would get better recognition result if we pass it to a French HOCR than to an English HOCR. This will have definite effect on the successive language identification accuracy.

The distinct contribution of this paper is to attempt character recognition based identification of language of a handwritten document. Languages covered by two different scripts namely Latin and Bengali are considered. For Latin script, English and French and for Bengali script, Bengali and Assamese languages are considered for the present experiment. For handwriting recognition, the state-of-the-art BLSTM-CTC [8] framework is used. Implementation of BLSTM-CTC based classifier for Bengali handwriting recognition is another significant by-product of this research work. For language identification, we use an off-the-shelf tool which is based on the analysis of statistical character N-gram distributions [9]. By default, it covers 49 languages for identification. Assamese is not in this list and hence language profile for Assamese was learned separately and added in the existing list of languages. Experimental protocol extensively investigates the effect of N-gram model used during recognition on the language identification accuracy. Influence of text length is also studied. Experience gathered from this work is discussed and future research perspectives are summarized.

II. The Approach

Our language identification method is schematically shown in Fig. 1. Consider there are $\mathcal{S}$ different scripts and $\mathcal{L}$ different languages. In our approach we at least need $\mathcal{S}$ different HOCR systems. Since HOCR systems make use of language model, they are more related to language than script and therefore, there is as such no HOCR particular to one script. Here we assume that if we have an HOCR for a language (Lj) trained using script (S_i) then the HOCR will be considered as the HOCR all the languages written in S_i and HOCR$_i$ will denote the HOCR associated to script S_i. For instance, if we have a French HOCR then it will be used to recognize any Latin based handwriting. Later on, we will check the effect if we train an HOCR using more than one language. Say, an HOCR is trained on both English and French and is used to recognize Latin-based handwriting.

As shown in Fig. 1, the handwritten document (D), written in only one script, is passed through a script recognizer in order to identify the script (S_i) of D. Next, the document is passed to HOCR$_i$ to recognize its content. The recognized output given by HOCR$_i$ is fed to the language identifier that identifies the language of D. Since a lot of research has already been done for script recognition, discussion related to script recognition is avoided here and we assume that script recognizer is available to us. The following two sub-sections discuss about the handwriting OCR systems and the language identifier used in our system.

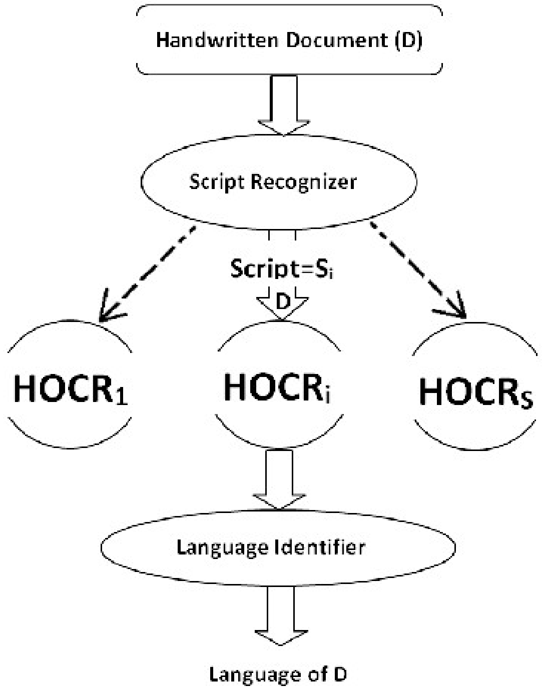

Fig. 1. Language Identification: the schematic diagram of our method.

A. BLSTM-CTC Based Handwriting Recognition

In order to transform a digital image into digital text that we can use for language identification we use an HOCR system. This system is based on a sliding window method extracting features that are then used as input to a BLSTM-CTC classifier.

The images are first preprocessed to reduce variability and noise. The preprocessing steps include : binarization, deslanting, deskewing and height normalization.

Following the preprocessing step we extract Histogram of Gradient [12] (HOG) features. The sliding window is 8px wide and 1px apart from the next position. The HOG features are computed on a grid of 2×4 sub-windows from which histograms are computed. Each histogram is composed of 8 directions. Hence each window is represented by an $8 \times 8 = 64$ dimensional feature vector. The sequence of feature vector of an image is then processed by the BLSTM-CTC.

The BLSTM-CTC is a novel neural network architecture that enables to compute character posterior probabilities on a sequence. It is composed of two recurrent neural networks with Long Short Term Memory (LSTM) neurons [13]. One recurrent neural network processes the signal chronologically while the other one processes the signal ante-chronologically. At every time step they hand over the information they have accumulated over time to the Connectionist Temporal

Classification (CTC) Layer[14] . This layer is a Softmax neural network layer that outputs character probabilities. It has the ability to avoid outputting ambiguous decisions, thanks to an additional 'blank' output. The BLSTM-CTC trained on the different scripts are composed of two BLSTM hidden layers with around 100 to 200 neurons on each layer. The CTC layer is composed of 90 outputs for Latin and 900 for Bengali (outputs account for the different character classes). For the two scripts we only considered the characters as classes, all punctuations and Unicode symbols are considered as noise and represented by the blank output of the CTC.

Bengali characters can be composed of up to four consonants and a single vowel, hence proving a potentially very large number of characters (up to 4000 different characters). In order to reduce this number we performed a preprocessing of characters to reduce the number of classes. Some vowels can be vertically split from the consonants, we decided to keep them as a separate class, hence enabling us to reduce the number of classes. Some character combinations are impossible, therefore they were removed. Thanks to this preprocessing we could reduce the number of classes to about 900. However it has to be stressed that not all characters were represented inside the training dataset, around a hundred characters are present in the database. The character sequences produced by the BLSTM-CTC are then used for language identification.

B. *Statistical Language Identifier*

We use a statistical language identifier where languages are treated as individual categories (English, Japanese, Hindi, etc.). A naive Bayes classifier is used to classify a piece of text into a language category. The classification is done by updating the posterior probabilities of categories by feature probabilities in each category as follows.

$$p(C_k|X)^{m+1} \propto p(C_k|X)^m \, . \, p(X_i|C_k),$$

where C_k denotes the k-th category, X be the input text, X_i be the feature of the text and m denotes the iteration number. The detection process terminates when the maximum probability is over 0.99999. Character n-gram (more specifically, Unicode's codepoint n-gram) are used as features for language detection.

A Java based open source (Apache License 2.0) implementation of the above model achieves over 99% precision for detection of 49 major languages [9] where text length is at least 200 characters. For a given piece of text, the tool returns the candidates and their probabilities. Though, by default, 49 major languages are supported but one can generate new language profiles from a training corpus. For example, Assamese is not in the list of 49 languages and therefore, for our purpose, we generated the language profile for Assamese.

III. Experimental Protocol

In this experiment, we have considered two scripts namely, Latin and Bengali and for each script we have considered two languages, English and French for Latin script and Bengali and Assamese for Bengali Script. Though English and French share the same script but variations in use of accent symbols is significant. Bengali and Assamese also do not share exactly the same script rather they vary in just two character shapes.

A. *Details of the Datasets*

For English and French data we have used samples from the MAURDOR database [10]. The database contains handwriting samples for three languages namely, French, English and Arabic. As the former two share the same script we have considered only French and English samples. In total, $20,000$ and $8,700$ lines are considered for French and English. Lines may contain a variable number of words and altogether there are $53,000$ and $20,000$ words for French and English, respectively. The dataset was divided following a ratio 60% for training data, 20% for validation data and 20% for test data.

In order to evaluate the language modeling ability of the BLSTM we designed three different training datasets for the latin alphabet. The first one is composed of french lines only, the second one of english lines only and the third one of english and french lines.

The Bengali dataset consists of about $2,300$ lines (about $21,000$ words) was recently prepared at the Indian Statistical Institute. The dataset reported in [11] has been used to get about $1,000$ images of handwritten Bengali lines and the remaining images have been collected in-house. For each text line image, groundtruthed text was produced manually.

B. *Experimental Results*

Before presenting the success rate of language identification, we first present the performance of the BLSTM-CTC based handwriting OCR (HOCR) systems. Next, the language identification accuracies under different experimental setups are presented.

1) Performance of the BLSTM-CTC based HOCRs: The latin model results are presented in table I, these results are measured on an all latin test datasets (containing French and English).

TABLE I. Results of the different latin models

Model	Character-level error rate	Deletions	Substitutions	Insertions	Sequence-level error rate
English	46.37%	9.71%	34.43%	2.23%	90.45%
French	35.22%	5.32%	25.25%	5.65%	84.00%
Latin	30.69%	7.87%	21.47%	1.35%	80.69%

The difference between the French and English results can be explained by the fact that the English HOCR was trained with less data than the French HOCR.

Bengali HOCR character error rate is 24.60%. Among these errors, the substitution, deletion and insertion errors are 18.91%, deletions 4.69% and 0.98%, respectively. The line (text lines which, on average, contain 9 words) level error rate is 99.40%. No post-processing was used on the output of the BLSTM-CTC.

2) Accuracy of Language Identification: As discussed before, we have three HOCRs in hand, two (e.g. French and English HOCRs) for Latin-based languages and one, i.e., Bengali HOCR for Bengali-script based languages. Let us first assume that for identifying Latin-based languages we will use any of the two HOCRs assuming that the HOCR will recognize both French and English with "acceptable" accuracy so that language identification would not be affected.

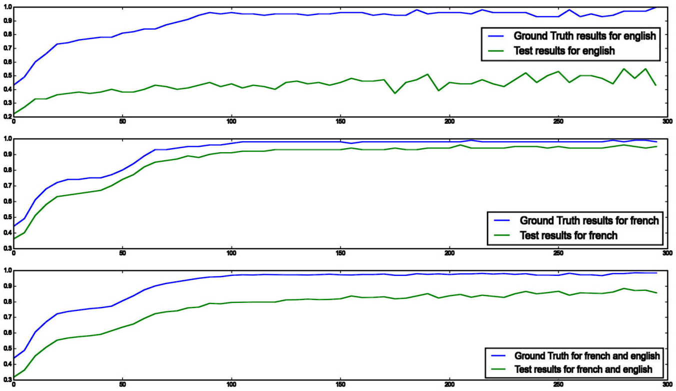

Fig. 2. Latin-based Language Identification: French trained HOCR.

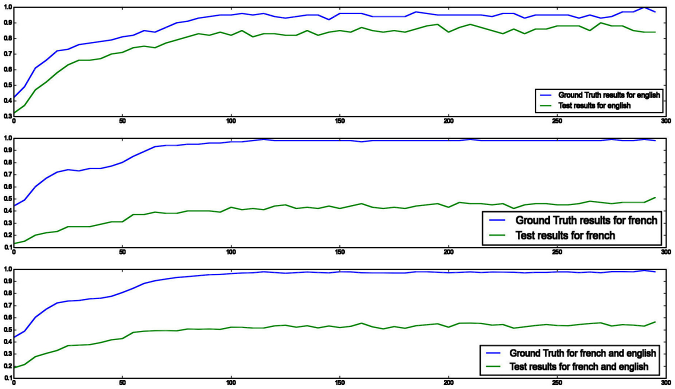

Fig. 3. Latin-based Language Identification: English trained HOCR.

The Fig. 2 presents the results for identifying language of a document handwritten in English or French when the French HOCR is used for handwriting recognition. For latin script, the classification is achieved among more than 30 languages. The text length in terms of number of characters is plotted along the x-axis. The blue lines show the language identification accuracy if the corresponding groundtruthed text (which are available for each handwritten line) is sent to the language identifier. The upper block of Fig. 2 shows the performance curve for identifying English, the middle is for identifying French and the lower block shows the performance for identifying language in English or French handwriting. It is to be noted that accuracy is low when an English handwriting is passed through the French HOCR.

Almost similar trend (Fig. 3) is observed when we replace the French HOCR by the English HOCR. It shows good language identification accuracy when English handwriting is processed but poor accuracy for French handwriting. Next, we trained the HOCR using the training samples for both French and English data. When this newly trained HOCR is used as handwriting recognizer, the language identification accuracies are shown in Fig. 4. The upper and middle blocks show that unlike in Figures 2 and 3 accuracy does not fall for any particular language and therefore, the overall accuracy (lower block) is also improved.

Results for identifying Bengali script-based languages (i.e. Bengali and Assamese) are shown in Fig. 5. Language identification accuracy for documents handwritten in Bengali is very impressive (the middle block) but it is extremely poor for Assamese handwriting. The overall success rate is about 50% but it is mainly due to the accuracy for correctly identifying Bengali language. Note that the HOCR used in this experiment is also trained using handwriting in Bengali language. We could not find enough handwriting in Assamese language and hence, retraining of the HOCR on Assamese data was not possible.

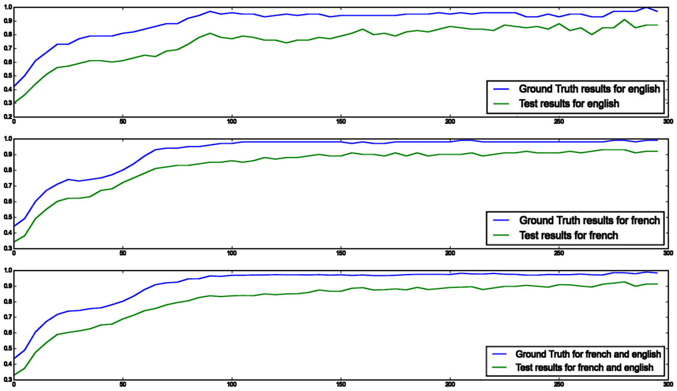

Fig. 4. Latin-based Language Identification: French and English trained HOCR.

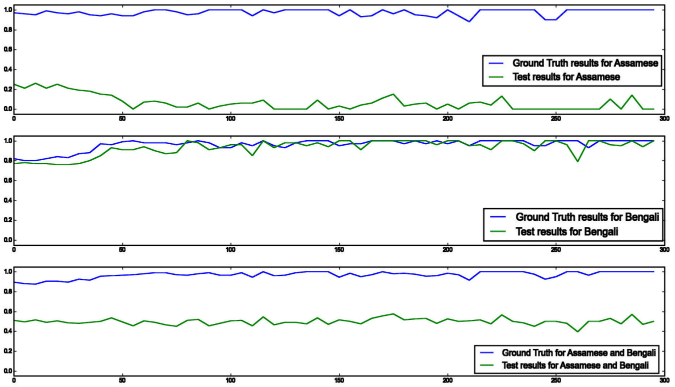

Fig. 5. Bengali Script-based Language Identification: Bengali trained HOCR.

3) Discussion: The results presented before show several important aspects of language identification through character recognition. As the BLSTM-CTC based recognition scheme learns language structure from the training data, the HOCRs cannot be treated as script-specific. They are rather language dependent HOCRs. This is the major reason why we get less accuracy in identifying English when French HOCR was used or vice versa. This effect is nicely amplified for Bengali/Assamese experiment. Though both languages share the same common script, their language structures (spelling pattern, use of words in a sentence, etc.) are very different from one another. This aspect has been nicely demonstrated by the poor accuracy for identifying Assamese language and the very high accuracy for recognizing Bengali language.

The language identification score is around or above 80% (for text length of more than 100 characters) when the same language HOCR is used. This is true for all the three languages (French, English and Bengali). However, whenever other language HOCR is used the language accuracy falls down to

40% or less (e.g. French HOCR, English language document or vice versa). However, when the HOCR training set covers samples from all the languages (e.g. French and English both), it behaves nicely in identifying any language. This has been tested with only two languages, the generalization of this observation requires experiments involving more languages.

Though the majority of the errors can be attributed to the performance of BLSTM-CTC based HOCR but sometimes errors come due to noise cleaning module of the statistical language identifier used in this experiment. In the noise filtering part, the language identification tool removes all-capital words, acronyms (e.g. UNESCO), place names (New York), punctuation, etc. as these do not represent language specific features. Therefore, if a document (which is a handwritten line in the present experiment) contains many such words and a few language-specific words then language identification becomes difficult. This is also well reflected in the language identification accuracy using groundtruthed data (the accuracy is not always more than 99% for all text lines).

IV. Conclusion

This work presents a novel method for language identification from handwritten documents. This area remained unexplored though the DIA community has seen many research efforts for script identification. The little amount of work that dealt with language identification before had also viewed the problem from script identification viewpoint and hence, glyph, connected-component or in general shape based features were only explored for language identification too. We presented a character recognition based approach with a view that a language is more related to the use of characters rather than their geometric shapes which is surely important for script identification.

The results clearly show that BLSTM-CTC based handwriting recognition is language dependent within the same script. Design of a less dependent character recognizer may help in getting better language identification accuracy, but conversely it may get lower character recognition performance which may adversely affect the language recognition performance. Therefore, more research and experiment is needed to improve this trade-off. It is also noted that as BLSTM-CTC embeds a character language model, it is largely dependant of the data set it was trained on. This remark not only concerns language but also the lexicon. If the learning lexicon is not representative of the whole lexicon of the language then we can expect having a biased BLSTM-CTC. Additional research is needed to explore this issue further.

One definitely significant exploration is the training of BLSTM-CTC on both English and French handwriting samples. This framework gives good accuracy for language identification. However, this has been tested only using two languages and this has to be extended for some more languages (e.g. German, Spanish, Italian, etc.) to claim this feature in general.

The present research has two more important byproducts. A Bengali handwriting dataset containing about $2,300$ handwritten lines along with their groundtruth is now available for doing further research on Bengali handwriting recognition. Earlier datasets (e.g. one in [11]) either do not contain line wise data, annotations, or are not available in public domain. This forced us to take up this initiative to create an freely available dataset for Bengali unconstrained handwriting recognition.

Development of a BLSTM-CTC based Bengali handwriting recognizer is the second byproduct of this work. The recognizer is giving about 75% accuracy in character-level recognition and almost 0% for line level (one line is consisting of about 10 words). One reason behind getting such poor accuracy can be attributed to i) the large number classes for which no BLSTM-CTC based HOCR has been developed before, and ii) the limited amount of data. Further research is needed to improve this score by manifold either by reducing the number of classes (if this number is hampering the classifier at all) or by employing suitable post-processing technique. Being slightly out of scope for the present problem, development of BLSTM-CTC based Bengali character recognizer has been present here in a very brief manner. We would like to report this experiment in another future communication.

As further extension of the current research on language identification, direct comparison between a shape feature based approach and the present approach could be the next task we want to take up. Moreover, we want to add more scripts and more languages in this experiment. There are many other Latin based languages that will be considered in the future extension of this study. At the same time some more scripts which are used to write many languages (e.g. Devanagari is used to write many languages) will be considered for generalization of the present approach.

References

[1] D. Ghosh, T. Dube, and A. P. Shivaprasad, *Script recognition–A review*, IEEE Trans. on PAMI, 32(12): 2142-2161, 2010.

[2] P. B. Pati and A. G. Ramakrishnan, *Word level multi-script identification*, Pattern Recognition Letters, 29(9): 1218-1229, 2008.

[3] G. Zhu, X. Yu, Y. Li, and D. Doermann, *Language identification for handwritten document images using a shape codebook*, Pattern Recognition, 42(12): 3184-3191, 2009.

[4] A. L. Spitz, *Determination of the script and language content of document images*, IEEE Trans. on PAMI, 19(3): 235-245, 1997.

[5] Y. H. Liu, C. C. Lin, and F. Chang, *Language identification of character images using machine learning techniques*, In ICDAR, 630-634, 2005.

[6] J. Hochberg, K. Bowers, M. Cannon, P. Kelly, *Script and language identification for handwritten document images*, IJDAR, 1999, 45-52

[7] T. Dunning, *Statistical identification of language*, published by Computing Research Laboratory, New Mexico State University, 1994.

[8] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, *A Novel Connectionist System for Unconstrained Handwriting Recognition*, IEEE Trans. on PAMI 31(5): 855-868 (2009)

[9] S. Nakatani, *Language Detection Library for Java*, http://code.google.com/p/language-detection/, 2010.

[10] MAURDOR campaign website, http://www.maurdor-campaign.org/

[11] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, *CMATERdb1: a database of unconstrained handwritten Bangla and Bangla-English mixed script document image*, in IJDAR 15(1):71-83, 2012.

[12] N. Dalal, B. Triggs, *Histograms of Oriented Gradients for Human Detection*, IEEE CVPR, 2005, 886–893

[13] S. Hochreiter, J. Schmidhuber, *Long Short Term Memory*, Neural Computation, 1997, 1735–1780

[14] A. Graves, F. Gomez, *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks*, International Conference on Machine Learning, 2006