

Conditional Random Fields for Online Handwriting Recognition

T.M.T Do and T. Artières

LIP6, Université Paris 6

8 rue du capitaine Scott

75015, France

Trinh-Minh-Tri.Do@lip6.fr, Thierry.Artieres@lip6.fr

Abstract

In this work, we present a conditional model for online handwriting recognition. Our approach is based on Conditional Random Fields (CRFs), a probabilistic discriminant model that has been generally used up to now in particular settings, for labeling and parsing of sequential data such as text documents and biological sequences. We propose to adapt these models in order to build systems for handwriting recognition. We propose a few systems whose architecture allows dealing with multimodal classes and exploiting segmental features that are much adapted to signal data like online handwriting.

1. Introduction

Classification, segmentation and labeling of sequential data are major problems for many application fields such as bioinformatics, on-line handwriting recognition, information extraction and so on. One of the main problems in these fields consists in transforming an observed sequence (e.g. a handwritten signal) into a sequence of labels. This task can be carried out at different levels. For instance one may seek to segment a handwritten signal that corresponds to a sentence in a sequence of words, characters and so on.

For decades, Hidden Markov Models (HMMs) have been the most popular approach for dealing with sequential data, e.g. for segmentation and classification, although they rely on strong independence assumptions and despite they are learned using Maximum Likelihood Estimation which is a non discriminant criterion. This latter point comes from the fact that HMMs are generative models and that they define a joint probability distribution on the sequence of observations X and the associated label sequence Y .

Various methods were proposed to build on Markovian models while first relaxing independence assumptions and second introducing discriminative information. Hence a number of segmental models were proposed that aim at handling correlation and dependency between successive observations. Extensions of Markovian models have been proposed that rely on (e.g. polynomial) trajectory models in each state or on autoregressive models [8], [10], [15]. Besides

a number of methods have been proposed to learn discriminant systems based on Markovian models or more generally on generative models [3], [9], [14]. Most often these works rely on kernel methods and exploit the framework of Support Vector Machines. Kernels are used to transform the data from a variable length representation space (e.g. variable length sequences) into a fixed dimension representation space. By doing so, one can then rely on any discriminant model such as Support Vector Machines that are adapted for input data in a fixed dimension representation space.

An alternative has been proposed recently as a few probabilistic conditional models have been investigated for replacing traditional HMMs in sequence processing tasks. These models are conditional models which attempt to model directly the conditional probability distribution $P(Y/X)$. Hence, these models do not require modeling observation sequences X hence no simplifying assumptions over X are necessary. Among these models we can cite Maximum Entropy Markov Models [1] and Conditional Random Fields [5] that are the most popular ones. These models rely on the definition by hand of a set of (eventually many) features, computed from observations X and label sequence Y , and the optimal learning of a classifier where each label sequence is viewed as a possible class for an input observation sequence. Up to now these models have been used in particular settings, conditional models have been first proposed and then mainly investigated for dealing with text documents to perform information extraction, FAQ segmentation, POS-tagging etc. As a consequence, these models have mainly been used with textual data for which features may be efficiently hand defined. Also traditional training algorithms for Conditional Random Fields (CRFs) require completely labeled training sequences that are available for information extraction tasks but are usually not available for signal classification tasks.

This contribution investigates the use of conditional models for the classification of continuous signals such as speech and online handwritten signals. This requires a few adaptations. Firstly in signal processing tasks input data are traditionally sequences of feature vectors in R^D and are then much different from textual document representations. Secondly, isolated observations are often much less informative than a segment of few

successive observations. For instance in on-line handwriting, a point coordinate is not much informative while a segment of ten points coordinates correspond to a stroke in a character whose configuration is worth modeling. Hence segmental features are generally much more informative and discriminant in signal recognition tasks. Thirdly, classes are naturally multimodal, for instance there are several ways writing style for a character, e.g. “b”. Lastly, training data are usually not completely labeled in signal processing tasks. The class of an observation sequence (e.g. it is an “a”) is known but the complete state (i.e. label) sequence is not available. Training algorithms should then be adapted to this partial labeling of training data.

The paper is organized as follows. We start by introducing Conditional Random Fields. Then we describe CRF architecture for dealing with multimodal classes and describe training algorithms for learning these models with partially labeled data. We also show how we exploit segmental features for online handwriting signal. At last, we provide experimental results on on-line handwritten character recognition, where we compare our conditional models with more standard Markovian models.

2. Conditional Random Fields

Sequence labeling consists in identifying the sequence of labels $\hat{Y} = y_1, \dots, y_T$ that best matches an observation sequence $X = x_1 \dots x_T$:

$$\hat{Y} = \arg \max_Y P(Y/X) \quad (1)$$

CRFs are a particular instance of Markov random fields. Figure 1 illustrates the difference between traditional HMMs and CRFs where they are both represented as graphical models. HMMs (Figure 1-a) are directed models where independence assumptions between two variables are expressed by the absence of edges. The probability of the state at time t depends only on the state at time $t-1$, and the observation generated at time t only depends on the state of the model at time t . The probability distribution may be factorized using these independence assumptions. CRFs are undirected graphical models, Figure 1-b shows a CRF with a chain structure. The probability distribution for such a model may be expressed as a product of potential functions, one per clique of the graph. With the chain structure in the figure, the random variable y_t depends on its neighbours (y_{t-1}, y_{t+1}, X) . One must notice that CRFs being conditional models, node X is observed so that X has not to be modeled. Hence CRFs do not require any assumption about the distribution of X . From the random field theory, one can show [5] that the likelihood $P(Y/X)$ may be parameterized as:

$$P(Y/X, W) = \frac{e^{W \cdot F(X, Y)}}{\sum_{Y'} e^{W \cdot F(X, Y')}} = \frac{e^{W \cdot F(X, Y)}}{Z_W(X)} \quad (2)$$

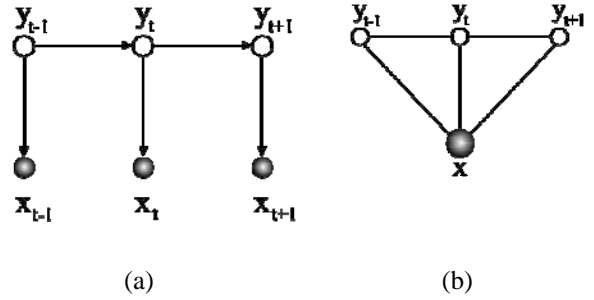


Figure 1. Dynamic representation of an HMM (a) and a CRF with a chain structure (b) as graphical models, where grey nodes represent observed variables.

where $Z_W(X) = \sum_{Y'} e^{W \cdot F(X, Y')}$ is a normalization factor and Y' ranges over all allowed segmentations (i.e. sequence of states), $F(X, Y)$ is a feature vector and W is a weight vector consisting of the model’s parameters. Features $F(X, Y)$ are computed on maximal cliques of the graph. In the case of a chain structure (Figure 1-b), these cliques are edges (y_{t-1}, y_t) and vertices (y_t) .

In some cases there is a need to relax the Markovian hypothesis by allowing the process not to be Markovian within a state. To do this [12] proposed semi-Markov CRFs (SCRFs). The main idea of these models is to use segmental features, computed on a segment of observations associated to a same node (i.e. state). Consider the segmentation of an input sequence $X = x_1, x_2, \dots, x_T$, this segmentation may be described as a sequence of segments $S = s_1, s_2, \dots, s_J$, where $J \leq T$ and $s_j = (e_j, l_j, y_j)$ where e_j and l_j stand for the entering and leaving times in state y_j . Segmental features are computed over segments of observations x_{e_j}, \dots, x_{l_j} corresponding to a particular state y_j . SCRFs aim at computing $P(S/X)$ defined similarly as in Eq. 2. To enable efficient dynamic programming, one assumes that the features can be expressed in terms of functions of X , s_j and y_{j-1} , these are segmental features:

$$F(X, S) = \sum_{j=1}^{|S|} F(X, y_j, y_{j-1}, e_j, l_j) \quad (3)$$

Inference in CRFs and SCRFs is performed with dynamic programming like algorithm. Depending on the underlying structure (chain, tree, or anything else) one can use the Viterbi algorithm, Belief Propagation or Loopy Belief Propagation [6], [16].

Training a CRF consists in maximizing the log-likelihood $L(W)$ based on a fully labeled database of K samples, $\{(X_k, Y_k)\}_{k=1..K}$, where X_k is a sequence of

observations and Y_k is the corresponding sequence of states (i.e. labels).

$$L(W) = \sum_{k=1}^K \log P(Y_k / X_k, W) \quad (4)$$

This criterion is convex and is maximized using gradient ascent methods. Note that computing $Z_w(X)$ includes a summation over an exponential number of label sequences that may be computed efficiently using dynamic programming. Training a SCRf is very similar to CRF training and also relies on a fully labeled database.

3. Conditional Random Fields for on-line handwriting recognition.

3.1. Sequence classification with CRFs

CRFs, as we already mentioned, are traditionally used for labeling sequential data. To build a system based on CRFs for sequence classification we investigated the use of a CRF whose architecture is based on a structure of chain for each class (hereafter named *chained CRF*). A chain structure is indeed natural for modeling sequences of the same class (e.g. writing samples of a character). Each state in the chain structure is used to model a specific part of the samples. For instance for on-line handwriting signals, the successive states of a model for character "a" correspond to the beginning of the writing of character "a", the medium part and so on. The architecture of the system is illustrated in Figure 2. It is based on a chain structure, where there is one chain per class. The chains may have different number of states, depending on the complexity of the character being modeled. Of course, the "dynamic" representation of this model is similar to Figure 1-b, the difference here comes from the fact that all the transitions between nodes are not authorized. Learning such a model results in a set of parameters W that allow discriminating between observation sequences of all classes.

In signal classification and segmentation tasks labeling of training data is usually often tiny and limited to the class label. The main reason is that there is no known semantic associated to states before learning so that complete state path cannot be manually determined. Despite training algorithms for CRFs usually require this label information, we describe here how to learn a CRF without a complete labeling of training observation sequences. This is done by introducing hidden variables for the state sequences.

Consider a training data set of K samples, $\{(X_k, Y_k)\}_{k=1..K}$, where X_k is a sequence of observations and Y_k is the class label corresponding to X_k . Then,

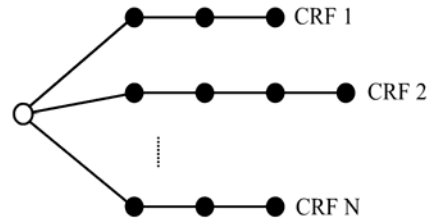


Figure 2. Mixture of chained CRF (i.e. with a chain structure) for sequence classification. Each chain corresponds to a class.

the conditional probability $P(Y_k / X_k)$ may be written as follow:

$$P(Y_k / X_k) = \sum_{S \in S(Y_k)} P(Y_k, S / X_k) = \frac{\sum_{S \in S(Y_k)} e^{W.F(S, Y_k, X_k)}}{Z_w(X_k)} \quad (5)$$

where $Z_w(X_k)$ is a normalization factor, S indicates a segmentation for X_k (i.e. a sequence of states), $S(Y_k)$ indicates the set of possible segmentation (sequence of states) for a sequence whose class is Y_k . Here, $S(Y_k)$ corresponds to the set of all sequences of states in the chained CRF corresponding to class Y_k . This modeling is similar to a mixture of models [2]. Unfortunately although the normalization $Z_w(X_k)$ may be computed using a forward-backward like dynamic programming routine it comes with numerical problems. To simplify the implementation we chose to approximate the preceding quantity by:

$$P(Y_k / X_k) = \sum_S P(Y_k, S / X_k) \approx \frac{\max_S e^{W.F(S, Y_k, X_k)}}{Z'_W(X_k)} \quad (6)$$

where the new normalization factor is defined as $Z'_W(X_k) = \sum_Y \max_S e^{W.F(S, Y, X_k)}$, which may be computed using a Viterbi-like dynamic programming algorithm.

Training aims at estimating parameters W that maximize the conditional log likelihood $L(W)$:

$$L(W) = \sum_{k=1}^K \log P(Y_k / X_k, W) \quad (7)$$

with :

$$\log P(Y_k / X_k, W) = e^{W.F(\hat{S}_k^Y, Y_k, X_k)} - \log Z'_W(X_k) \quad (8)$$

where $\hat{S}_k^Y = \arg \max_S e^{W.F(S, Y, X_k)}$ stands for the best segmentation of sample X_k given the class Y_k .

The learning algorithm consists of two iterative steps, segmentation and maximization. In the

segmentation step, we use a segmental Viterbi algorithm to find the best state sequence for each class given the signal and the current parameters. Based on these segmentations, we update the model parameters to increase training data likelihood through gradient ascent method:

$$W \leftarrow W + \varphi \frac{\partial L(W)}{\partial W} \quad (9)$$

It should be noted that because of hidden variables (segmentation paths) the criterion may have local maxima, so that optimization is not warranted to find the global optimum.

3.2. Handling multimodality

In signal processing applications classes are often multimodal. For instance in the case of on-line handwriting, there are multiple allograph corresponding to various writing styles for a character. The system presented above in section 3.1 cannot handle this variability. To deal with multimodality we propose to extend this previous model by using a mixture of chained CRFs per class. The idea is that each chain of a mixture model is dedicated to an allograph of the character and should specialize during training. This system is illustrated in Figure 2 where there are M chained CRF for every class (character). These CRFs may be viewed as ‘‘allograph CRF’’ in the following. The number of allograph CRFs per class (M) may be used to tune the complexity of the system and will be called the model size of character models.

Training such a system raises the same kind of problem as in section 3.1, which is related to the absence of labeling information for training samples. The only available labeling information for a training sample is its class. Besides, the allograph corresponding to this training sample, together with the segmentation of this sample in the corresponding allograph CRF, are unknown. To deal with this we introduced a second kind of hidden variables, which are indicators of the allograph corresponding to training samples. Then:

$$P(Y_k / X_k) = \frac{\sum_{B \in B(Y_k)} \left[\sum_{S \in S(Y_k)} P(Y_k, B, S / X_k) \right]}{\sum_{B \in B(Y_k)} \left[\sum_{S \in S(Y_k)} e^{W \cdot F(S, B, X_k)} \right]} \quad (10)$$

$$= \frac{\sum_{B \in B(Y_k)} \left[\sum_{S \in S(Y_k)} e^{W \cdot F(S, B, X_k)} \right]}{Z_W(X_k)}$$

where B indicates the allograph (with value between 1 and M), $B(Y_k)$ indicates the set of the chained CRFs that correspond to class Y_k . Here S , $S(Y_k)$ and $Z_W(X_k)$ have the same meaning as in previous section. To avoid numerical problems and to simplify implementation one can choose to approximate the preceding quantity by approximating the summation in

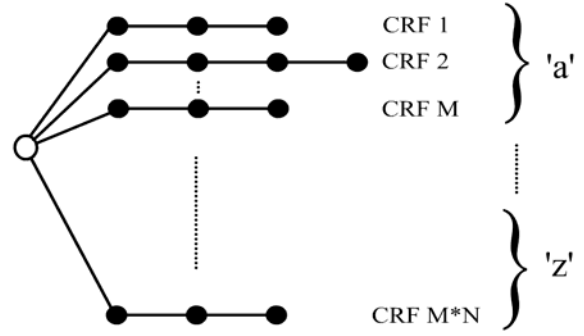


Figure 3. Architecture of a system for dealing with multimodal classes. There are one mixture of M allograph models per class, M is called the model size. Allograph models are segmental CRFs with a chain structure.

the numerator by a maximum, as has been discussed for Eq. 5 and Eq. 6.

4. Prior choices

Designing systems as those that we described in section 3 is not an easy task since many choices must be done before optimizing parameters.

4.1. Choosing system’s topology

One main choice concerns the topology of the system, the number of allograph CRFs per class as well as the number of states in an allograph model. This problem is a general one and concerns traditional Markovian models as well.

Various studies have been done to completely learn character models from the data, especially for HMM systems, e.g. [7], [13]. These latter works are based on the building of a HMM from only one training sample and exploit a representation of handwriting samples as sequences of direction codes. Here, we built on our previous work for designing the topology of CRF systems. Our system in [11] allows designing automatically the topology of Markovian models where each character is modeled with a mixture of allograph left-right Markovian models.

In this study the topology of CRF models (number of chained CRF allograph models and the number of states for each allograph model) is built using our previous technology. First a HMM system is built according to the method presented in [11]. Then we build a CRF system with the same topology which is then learned with the algorithm described in section 3. Moreover, we use in these CRF systems the same kind of features that are used in the HMM system, we describe these below.

4.2. Features

In CRF, the choice of efficient features is mandatory for reaching good performance, e.g. features should allow distinguishing between allograph of all characters. When dealing with on-line handwriting main features

are related to shape, duration and position of parts of the writing [11].

Given an observation sequence X , an allograph indicator B^i , a segmentation S , and a vector of weights W , the score used in Eq. (9) is defined as:

$$\text{score}(X, S, B^i, W) = e^{W^i \cdot F(B^i, S, X)} \quad (11)$$

where $F(S, B^i, X)$ is a feature vector whose features are computed for each state. These features are derived from the ones used in [11] and have been proved efficient for HMM-based systems. We used four types of features, each one is computed from a segment of observations that is associated to a state and an ideal segment corresponding to the state. We use a shape feature, a duration feature, a position feature and a continuity feature. For instance the shape feature measures the shape similarity between an observed segment and the ideal shape one should observe in the state.

In order to obtain the most flexible possible model, we used four feature functions per state instead of five feature functions for all the states. Then, the number of features is $\sum_{j=1}^{N*M} 4 * |B^j|$ where N is the number of classes, and M is the number of chained CRFs per class. $|B^j|$ is the number of states in the j^{th} allograph CRF.

5. Experimental results

We performed experiments on the international benchmark UNIPEN database [4]. We worked on parts of this database including signals corresponding to the 26 lowercase characters from 200 writers. Our database includes approximately 60000 samples from which 33% are used for training and 66% are used for testing. Samples are on-line signals, i.e. temporal signal consisting of successive coordinates of an electronic pen. Handwriting signals being much variable with many allographs the use of mixture models or systems based on prototypes is widespread. Among popular techniques, HMM have been shown to be powerful.

We report comparative results of the reference HMM system we discussed in section 4.1 and of our CRF systems. Table 1 reports accuracies for the classification of the twenty-six lowercase characters. Results are given as a function of the model size (i.e. the number of allograph CRFs per character), where we use the same model size for all characters. As may be seen, whatever the model size, CRF systems outperform the HMM system. Also, designing explicitly CRF systems for dealing with multimodal classes allows improving performance as it is the case for generative systems based on HMMs. Increasing the model complexity through increasing model size allows to improve both systems, HMMs and CRFs. It seems like HMM systems could reach the same performance as CRF systems provided enough allograph models are used per character. This is an open question however since this

would require additional experiments that we did not performed up to now. What we can say from these results, at least, is that CRF systems being discriminant systems require less parameters (i.e. a lower model size) to reach similar performance as generative HMM systems.

Table 2 gives more insight on the difference between generative and discriminant systems by focusing on a difficult classification problem with two classes. It report comparative results of the reference HMM system and of our CRF systems for the classification of character “g” and character “q”. As may be seen the difference between non discriminant HMMs and discriminant CRFs is much more significant here. For this couple of very confusable characters CRFs allows reaching high accuracy even with a low model size while the HMM system cannot reach such high accuracy even with increased model size. This superiority of our CRF systems was less clear in Table 1 where such performances for confusable characters were averaged when considering the twenty-six lowercase characters.

In order to get more information on how the discriminant training of CRFs works we explored what was learned in the models. Recall that in chained CRFs each state roughly corresponds to a part of the writing of a character, i.e. a stroke. Also, in our systems, as discussed in section 4.2, each state has its own features and weights. These weights can be thought of as indicators of the importance of features and states to discriminate between allographs and characters. In a way, the more discriminant is a state (i.e. the part of the writing that is modelled by a state) the bigger the weights are in that state.

To investigate this, we computed some statistics on the CRF system that has been learned for discriminating between the two confusing characters “g” and “q”. Then we computed the average of the weights in each state of each CRF model and plotted character samples with grey levels corresponding to the obtained values in each state. Figure 4 shows a typical result for two samples of these characters. One may see on these samples that the learned CRF models give more importance to the ending parts of the writings and give much less importance to the similar parts of the writings of the two characters (i.e. beginning of the writings). As expected, the learned CRFs have focused on what distinguishes these characters.

6. Conclusion

In this paper we investigated the adaptation of Conditional Random Fields for signals recognition tasks and especially for on-line handwritten signals. We investigated CRF architectures for dealing with multimodal data and proposed to use segmental features. We described training algorithms able to cope with partially labelled only training databases where labels consist in class information only. We provided experimental results showing that CRF systems significantly outperform our more traditional (state of

the art) Markovian system. We report isolated character recognition experiments but our systems may be extended to word recognition and segmentation tasks. This is one of our perspectives.

References

- [1] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation". *International Conference on Machine Learning*, 2000.
- [2] A. Quattoni, M. Collins and T. Darrel, "Conditional Random Fields for Object Recognition". In *Advances in Neural Information Processing Systems 17*, 2004.
- [3] C. Bahlmann, B. Haasdonk and H. Burkhardt, "On-line Handwriting Recognition using Support Vector Machines - A kernel approach". *International Workshop on Frontiers in Handwriting Recognition*, 2002.
- [4] I. Guyon , L. Schomaker , R. Plamondon , M. Liberman and S. Janet . "UNIPEN project of on-line data exchange and recognizer benchmark", *International Conference on Pattern Recognition*, 1994.
- [5] J. Lafferty, A. McCallum and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". *International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 2001, pp. 282–289.
- [6] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference". Morgan Kaufmann Eds., 1988.
- [7] J.J. Lee, J. Kim and J.H. Kim, "Data-driven design of HMM topology for on-line handwriting recognition". *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 15, n° 1, 2001, pp 107-121.
- [8] L. Deng, M. Aksmanovic, X. Sun and C.F.J Wu, "Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States", *IEEE Transactions on Speech and Audio Processing*, 1994, 2:4, pp507-520.
- [9] P.J. Moreno, P.P. Ho and N. Vasconcelos, "A Generative Model Based Kernel for SVM classification in Multimedia applications", In *Advances in Neural Information Processing Systems*, 2003
- [10] R. Chengalvarayan and L. Deng, "Trajectory Discrimination Using the Minimum Classification Error Learning". In *IEEE Transactions on Speech and Audio Processing*, 1998, 6:6, pp505-512 .
- [11] S. Marukatat, "Une approche générique pour la reconnaissance de signaux écrits en ligne". *Thèse de doctorat, Université Paris 6, LIP6*, 2004.
- [12] S. Sarawagi and W. Cohen, "Semi-Markov Conditional Random Fields for Information Extraction". *Advances in Neural Information Processing Systems*, 2004.
- [13] T. Artières T. and P. Gallinari, "Stroke level HMMs for on-line handwriting recognition", *International Workshop on Frontiers in Handwriting Recognition*, 2002.
- [14] T. Jaakkola , M. Diekhans and D. Haussler, "Using the Fisher kernel method to detect remote protein homologies", *International Conference on Intelligent Systems for Molecular Biology*, 1999.

Table 1. Accuracy of Hidden Markov Models and Conditional Random Fields for the recognition of 26 lowercase characters. Performances are shown as a function of the number of left-right models per character (i.e. model size).

Model Size	HMMs	CRFs
1	67.4%	76.4%
2	75.8%	82.7%
3	79.3%	84.6%
5	81.6%	85.9%

Table 2. Accuracy of Hidden Markov Models and Conditional Random Fields for the recognition of two lowercase characters 'g' and 'q'. Performances are shown as a function of the number of left-right models per character (i.e. model size).

Model Size	HMMs	CRFs
1	69.3%	88.0%
2	77.6%	94.8%
3	82.7%	95.2%
4	84.3%	95.3%
5	88.6%	95.6%

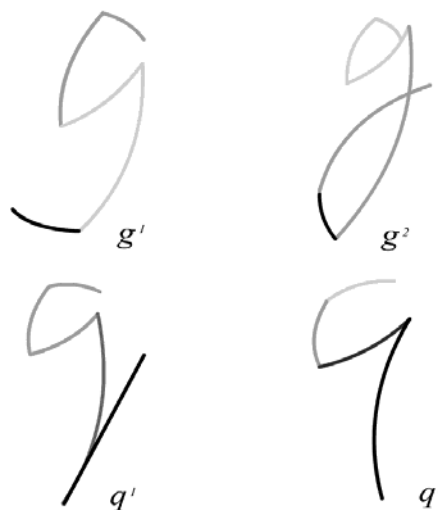


Figure 4. Samples of character "g" (top) and "q" (bottom) where grey level indicates the importance of the written part in a CRF system that has been trained to discriminate between the two characters.

- [15] Y. Ephraim, W.J.J. Roberts, "Revisiting autoregressive hidden Markov modeling of speech signals". In *IEEE Signal Processing Letters*, 2005, 12:2, pp166-169.
- [16] Y. Weiss Y, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology". *Neural Computation*, 13:2173-2200, 2001.