PhD Position 3 - Transliteration and Transcription of Historical Manuscripts

**Supervisor:** Prof M Cheriet  
**Location:** Synchromedia Lab, ETS, Montreal, Quebec, Canada  
**Starting Date:** September 1st, 2014  
**Keywords:** Transliteration, Image representation, Language model, HMM, Graphical Models.

**Objectives**  
The main goal of this research/proposal is to approach the challenge of transliteration and transcription of historical manuscripts, especially at the scale of collections of hundreds of manuscripts, in an objective, generalizable, and robust methodology. The main challenges of ancient documents are the use of different styles for the text appearance (visual level) and of a different grammar (linguistic level). Although many machines and also their associated features have been studied and used extensively along neural network approaches to extract visual features, there is a big room for data-driven methodological research in this area, in particular because of highly subjective nature of parameter selection (or, in a more general term, model selection) of a model design. The core of this PhD position/proposal is to develop a new model for local image data representation. It will be based on existing features, and combined with the representation learning framework in order to adapt the model to each manuscript. Sequential (linear) recognition models, such as HMM and recurrent neural network (RNN) will be used to generate the transliteration, using a language model (lexicon, N-gram). A secondary task will be to adapt a generic language model to a given manuscript, using the linguistic information contained in the manuscript itself. An advanced graphical model combining multiple source of linguistic information will be used for this purpose and integrated with the recognition model. As it can be noticed from the description, there is a tight correlation among PhD positions, Transliteration 1, Transliteration 2, Representation 2, and Data 1. Therefore, it is required that these students work in a close collaboration with each other all time.

**Research Questions**  
What would be a good image representation model for the task of transliteration? How to learn the representation model for the data at hand? Is it possible to integrate seamlessly the representation model with the recognition model (HMM, RNN)? What is the impact of the different language model on the transliteration? Is it possible to adapt the a given language model to fit a given manuscript from the information contained in the manuscript itself?

**Domains**  
Machine Learning; Concept Learning; Semi-Supervised Learning; Reinforcement Learning; Active Learning; Transfer Learning; Optical Character Recognition; Handwriting Recognition; Semantic Analysis; Context Modelling; Statistics; Graphical Models; Spatial Graphs; Graph Theory; Image Processing; Signal Processing; Mathematics; Scientific Programming;

**Requirements**  
A master in a relevant major; A CV of skills and experience in the relevant Domains; Possible published or under preparation papers relevant to the context of this PhD position.
References


[Cheriet2013] - Cheriet, Mohamed; Farrahi Moghaddam, Reza; Hedjam, Rachid; Visual language processing (VLP) of ancient manuscripts: Converting collections to windows on the past, GCC2013, 407-412, 2013. DOI: http://dx.doi.org/10.1109/IEEEGCC.2013.6705813

[Farrahi2012a] - Farrahi Moghaddam, Reza; Farrahi Moghaddam, Fereydoun; Cheriet, Mohamed; A new framework based on signature patches, micro registration, and sparse representation for optical text recognition, ISSPA2012, 1259-1265, 2012. DOI: http://dx.doi.org/10.1109/ISSPA.2012.6310485

[Chherawala2013] - Chherawala, Youssouf; Roy, Partha Pratim; Cheriet, Mohamed; Feature design for offline Arabic handwriting recognition: handcrafted vs automated?, ICDAR2013, 290-294, 2013. DOI: http://dx.doi.org/10.1109/ICDAR.2013.65


[Farrahi2011] - Farrahi Moghaddam, Reza; Cheriet, Mohamed; Beyond pixels and regions: A non-local patch means (NLPM) method for content-level restoration, enhancement, and reconstruction of degraded document images, Pattern Recognition, 44(2), 363-374, 2011. DOI: http://dx.doi.org/10.1016/j.patcog.2010.07.027