

INTERVIEW

**Christopher Kermorvant,
A2iA DocumentReader & Cursive
Research Department Manager**

Research remains at the heart of A2iA's developments. Today, A2iA operates one of the world's largest private research centres specialized in cursive handwriting recognition. This approach allows A2iA to offer highly innovative software that addresses today's changing market. Christopher Kermorvant tells us about his team's research on Arabic handwriting recognition.

When did you start working on Arabic recognition and why?

We first started working on Arabic recognition in 2006 as part of a CIFRE thesis. At the time, we were looking for an ambitious thesis topic that would allow A2iA's technology to move into new alphabets and that would fill a gap in the market. Recognition tools for printed characters already existed but there was nothing available for cursive handwriting. There was therefore a niche for A2iA.

What projects are you working on at the moment?

We are focusing our work on recognition competitions organized by the international scientific community. We recently took part in the Arabic handwriting recognition competition at ICDAR 2009 (International Conference on Document Analysis and Recognition). We placed first amongst companies and second overall (companies and university laboratories). At the same time, the defence and military intelligence sector is becoming increasingly interested in Arabic recognition, and we are currently working on a competition dedicated to this area. We are very interested in being part of this type of program; It is a good way to improve our technology and to incorporate these developments into A2iA's software to meet an increasing demand for document management from Middle Eastern companies and those looking to recognize Arabic handwriting throughout the world.

What is special about the Arabic language?

Arabic is made for A2iA, even printed words are cursive! All joking aside, the Arabic alphabet has 28 letters to which the different letter forms need to be added. One letter can in fact have 4 different forms depending on whether it is at the beginning, in the middle or at the end of a word. This makes 4 times 28 different characters.

In addition, some letters are only distinguished by the presence of one or more dots. There are also diacritical signs, like the accents in French, which are more numerous than in Latin

languages and some of which are not compulsory in writing. The reading direction, from right to left, really isn't the biggest problem! The real trouble lies in the complexity of the morphology of Arabic, i.e. in the construction of the words. This is a real obstacle, especially for non-mother tongue researchers. We must now expand the skills of our team to be able to understand the meaning of complete documents.

Can you explain in a few words how automated recognition of Arabic works?

The engine starts by "cleaning up" the images of the scanned documents, then identifying the lines and cutting them into words. Once a word has been identified, the engine uses two recognition techniques: division into "graphemes" (letters or parts of letters) and the "sliding window" technique. This second technique is fairly innovative: it combines various measurements (slope, pixels, loops, etc.) to recognise the word. In fact, the technology used is the same as for Latin languages. We just have to adapt it to the complexity of the Arabic language.

What is the outlook and when do you expect to see results?

At present, our engine automatically recognizes isolated handwritten words with a limited vocabulary. The short-term goal is to increase the size of the vocabulary and then recognize whole documents. Given the resources allocated to development of this technology, we expect to have the first prototypes ready by the end of the year.

CONTACT

EMEA : + 33(0) 1 44 42 00 80
AMERICAS : +1 (917) 237 0390
mail : contact@a2ia.com
web: www.a2ia.com

