

## **Extração automática de texto a partir de resíduos morfológicos com atributo em área**

**Cristina Almeida Magalhães**

### **Resumo**

O aumento vertiginoso da produção de conteúdos audiovisuais tem despertado o interesse dos operadores de televisão e outros produtores de conteúdos audiovisuais pela construção de bibliotecas digitais que armazenem arquivos de conteúdos multimídia para posterior reutilização. Para satisfazer esta necessidade são necessários sistemas capazes de tratar a informação audiovisual em termos de armazenamento, transmissão, procura e visualização. Desta forma, a existência de sistemas (automáticos e eficientes) de descrição, indexação e procura de conteúdos multimídia torna-se necessária e, para isso, vários métodos de processamento têm sido desenvolvidos ao longo dos últimos anos. Em muitos desses métodos, privilegia-se a informação textual existente nas imagens e nos vídeos, que, por ser uma fonte de informação com elevado valor semântico, torna possível, a produtores e consumidores, a indexação e busca de conteúdos de forma acessível e intuitiva. Os métodos desenvolvidos resolvem, em certas condições, o problema da extração de texto em imagens ou seqüências de vídeo, porém, ainda mantêm algumas limitações. As maiores dificuldades advêm da existência de caracteres com diferenças de tamanho, orientação, perspectiva, espaçamento, apresentação de cores na mesma linha ou palavra, fraco contraste em relação ao fundo e baixa resolução. A fim de melhorar o reconhecimento dos textos, foi realizado um processamento digital nas imagens antes da aplicação de um sistema de reconhecimento de caracteres. Tal processamento foi fundamentado na aplicação do conceito de resíduos morfológicos com atributo em área, que leva em consideração apenas o requisito de tamanho dos componentes conexos para realizar a filtragem. Em nossos experimentos, foi atingido um reconhecimento de 97% dos caracteres lingüísticos em um sistema OCR livre.

## Abstract

The increasing amount of audiovisual information has led television operators and audiovisual content producers to invest in the development of digital libraries which could store audiovisual information for future retrieval. The development of those libraries depends on systems that are able to treat audiovisual information as efficient as possible with respect to storage, transmission, searching and viewing. Therefore, those systems must be able to analyse, describe, filter and retrieve multimedia information. With this objective in mind, several methods have been developed in last few years. Many of these methods are based on the extraction of textual information in images and video sequences, since it is a source of highly semantic information and thus allows the filtering and searching of this information by producers and consumers in a more intuitive and natural way. However, there are certain constraints. The major problems occur due to the fact that characters may appear in different sizes, orientations, perspectives, letter spacing, and different colors in the same word or sentence, weak background contrast and low resolution. In order to improve the recognition of texts, we have developed a digital image processing technique, which must be applied before the use of an optical character recognition system (OCR). The proposed technique is based on morphological residues attributed to area, and takes in consideration only components sizes to process the filtering. In our experiments, we have achieved a 97% average recognition level of linguistics characters in a free OCR system.