Papyrus FreeForm®

Often below a group of objects of type Amount

Typically right of

Likely in the lower third of the page

FreeForm® Approach

Self-learning module especially for recognition and extraction of unsorted and unstructured documents

Product Description

Papyrus FreeForm® is a software module for automated recognition of scanned but unsorted business documents of unknown structure and layout. The system also recognizes unstructured or poorly structured documents with great reliability and is suited for processing any kind of business documents, such as commercial invoices, delivery notes, order forms, job applications, and many more. *FreeForm*® is based on the latest methods in pattern recognition and represents the most current standards in the fields of print analysis (OCR, ICR, Voting), associative databases, fuzzy logic and neural networks.

Mode of function
The Option *FreeForm*® module extracts from a document image the name of the document class and the field data that have been defined for this class. The concept allows for an N:M relationship between images and documents: a document can stretch over one or more pages, and likewise an image can contain several documents - similar to the way a newspaper page contains several articles.

January 1st, 2012

Mode of operation

FreeForm® processes scanned document pages in three steps:

Image Preprocessing
FreeForm® prepares the image data for optimal recognition results:
- Recognition of document alignment and autorotation (+/-90 degrees or 180 degrees)
- Deskewing
- Dirt removal and/or background removal

Document Classification
FreeForm® evaluates the image based on distinctive class traits and process-ing rules that were worked out during the training phase:
- Keywords at predefined positions or free in the text
- Text phrases at predefined positions or free in the text
- Graphics, line elements and background
- Colors at predefined positions, plus page format

Data Extraction
FreeForm® extracts and reads all neces-sary field data from the recognized document class. This is accomplished through evaluation of the following distinctive traits and rules:
- Absolute position
- Relative position through reference objects
- Pattern comparison (occurring patterns of defined synonymous- or data type objects)

Features

- Processing of documents of unknown format
- Classification and sorting of mixed document batches
- Combination of the methods of classic form reading with full text analysis of free text fields
- Smooth fit into a high performance and ergonomic production environ-ment with integrated verification capability
- Integrated definition and recogni-tion training environment, interac-tive setting of rules, parameters and zones with instant verification during training runs
- Statistic analysis of the processed document volume with option for using the results for system tuning

Definition and Training

Ergonomics - *FreeForm*® shields you from the complex and time-consuming absorption of cryptic definitions and rule sets. Training is intuitive and conveniently introduced via example documents.

Selection of relevant classification marks is effected during the design- and training stage:
- Automatically - on Auto-Learn mode, *FreeForm*® detects the relevant marks for best possible distinction of document classes in a given stock of images all by itself.
- By the qualified user (application designer) - graphic and simple selection of the relevant objects in an image via mouse click.

The handling of extraction rules, too, is facilitated by hands-on learning with example documents.

Control - "What gets measured gets done!" Consequently, the *FreeForm*® development environment comes with an integrated tool for constant statistic monitoring and analysis of the current training status. The user is informed immediately about the effects triggered by his definition steps. Problems and their causes are pointed out clearly. Everything important is in plain sight at any time - even big tasks can be accomplished easily.

Repository - *FreeForm*® features an object repository. Once-created definitions such as
- Synonyms, words or phrases
- Field types - simple or composite (e.g. address blocks, charts)
- Document classes and subclasses (types) are clearly displayed and administered and can be comfortably reused; new definitions being derived from proven existing ones.

This results in:
- Quick creation of applications
- Easy and efficient maintenance and upgrading of applications

Integration/supplementing products

The Option *FreeForm*® is optimized for use with Papyrus Capture. Through the ActiveX interface, this option can also be used by other applications and capture platforms that support the integration of ActiveX servers. The ISIS Papyrus raw format is documented (ASCII/UNICODE with data on image, zone and single character). The *FreeForm*® application designer is an autonomous application which is not needed during operational use of the document capture system.

Technical Prerequisites

Hardware
- Pentium IV or higher
- 512 MB RAM (1 GB recommended)
- per each Recognition-module there should be a Reco-server designated for better performance

Software
- *Papyrus Capture* base system

Dataformats

Supports a large number of file formats, for instance Text (DOS, ANSI, Unicode, UUEncode, MIME), Microsoft Word (Version 6.0 or higher), Microsoft Excel (Version 2.x or higher), HTML, RTF, WordPerfect (Version 6.0 or higher), Wordstar, Microsoft Works, Powerpoint, Lotus WordPro, Microsoft Outlook Mail Format, AmiPro.

Training

ISIS offers a special in-house training service for its clients, the *Papyrus Capture* Workshop. Please contact us for more details.