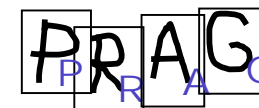


Image Spam Filtering by Detection of Adversarial Obfuscated Text

F. Roli, B. Biggio, G. Fumera, I. Pillai, R. Satta
Dept. of Electrical and Electronic Eng.
University of Cagliari, Italy



Pattern Recognition and Applications Group



- Faculty members

- F. Roli (group head)
- G. Giacinto
- G. Fumera
- L. Didaci
- G.L. Marcialis

7 PhD students

3 post docs

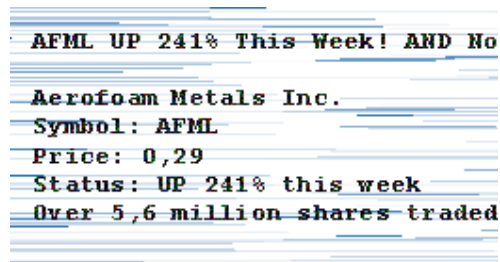
2 consultants

- Research interests

- Methodological issues
 - Multiple classifier systems
 - Classification reliability
- Main applications
 - Intrusion detection in computer networks
 - Multimedia document categorization, Spam filtering
 - Biometric authentication (fingerprint, face)
 - Content-based image retrieval

Evading spam filters with image spam

- Spam filtering: a pattern recognition task in adversarial environment
- Image spam
 - embedding text into images to evade text-based filtering modules
 - obfuscating images to evade OCR tools

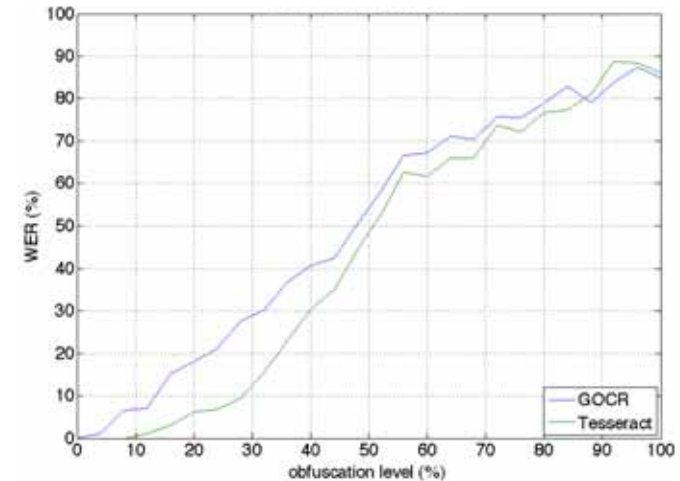


AFML UP 241% This Week! AND No
Aerofoam Metals Inc.
Symbol: AFML
Price: 0,29
Status: UP 241% this week
Over 5,6 million shares traded

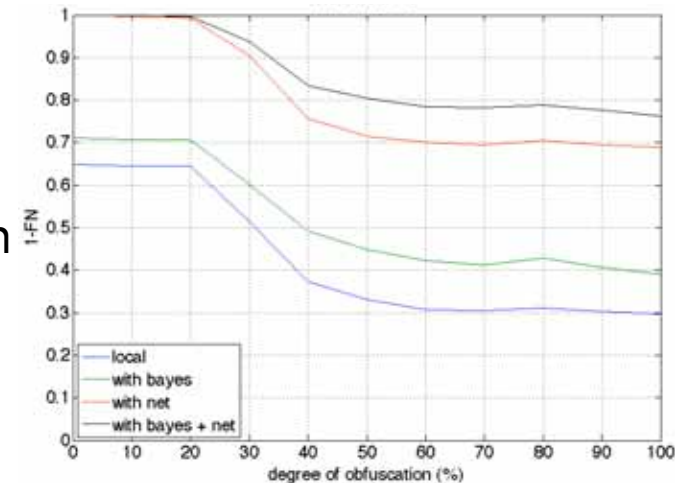


AFML UP 241% This Week! AND No
Stopping!
Aerofoam Metals Inc.
Symbol: AFML

- Our first contribution: a systematic evaluation of OCR and SpamAssassin performance through automatic generation of spam images with different degrees of obfuscation
- ⇒ Filtering adversarial obscured images is very difficult, if spammers' actions for evading classifiers are not taken into account explicitly



OCR performance



SpamAssassin performance

Image Spam Filtering by Detection of Adversarial Obfuscated Text

- Approaches alternative to OCR: image classification techniques
Aradhye et al., ICDAR 2005; Drezde et al., CEAS 2007
Drawback: “generic” image features (color distribution, etc.)
- Our approach: taking into account explicitly the adversarial environment by using features aimed at detecting a *specific* characteristics of image spam: the presence of obfuscated text
- Adding our features to “generic” ones improves the ham/spam discriminant capability of an image classifier
 - 2,006 real spam images (personal)
 - 3,297 real ham images (Drezde et al., CEAS 2007)
 - “generic” features: Aradhye et al., ICDAR 2005

