Classification Using Various Machine Learning Methods and Combinations of Key-Phrases and Visual Features

Yaakov HaCohen-Kerner[1], Asaf Sabag[1,]
Dimitris Liparas[2], Anastasia Moumtzidou[2], Stefanos Vrochidis[2], Ioannis Kompatsiaris[2]

[1] Dept. of Computer Science, Jerusalem College of Technology – Lev Academic Center, Jerusalem, Israel

[2] Centre for Research and Technology Hellas, Information Technologies Institute, Thermi-Thessaloniki, Greece

The first International KEYSTONE Conference (IKC-15),
8-9 September 2015,
Coimbra, Portugal

COST Action IC1302: KEYSTONE –
semantic KEYword-based Search on sTructured data sOurcEs
&
COST Action IC1307: iV&L – Vision-Language Integration Models

Motivation

- Challenge: Automatic and error-free classification of news documents into a set of categories
- In many cases, news documents contain multimodal information including
  - Textual descriptions
  - Images
- Most of the approaches consider only textual data for classification
- Investigate whether the combination of visual and textual features can improve the classification accuracy

Category-based classification of news documents

- Combination of visual and textual features (2 modalities)

- Comparative classification study
- 3 general types of features
  - Textual N-gram features
  - Visual low-level features
  - Textual N-gram + Visual features
- 3 different supervised machine learning (ML) methods
  - J48 decision tree algorithm
  - Random Forests (RF)
  - Sequential Minimal Optimization (SMO)

Related work (1/3)

Document classification

The majority of the studies make use of textual features

- Gamon et al. (2008)
  - Maximum entropy classifier on unigram features
  - Detect emotional charge around links to news articles in posts from political weblogs
- Swezey et al. (2012)
  - Approach for identifying and classifying contents of interest related to geographic communities from news articles streams
  - Bayesian text classifier
- Tang et al. (2014)
  - Learning of sentiment-specific word embedding for twitter sentiment classification
  - Embedding for unigrams, bigrams and trigrams separately using three developed neural networks

Related work (2/3)

Approaches that use only visual features

- Shin et al. (2001)
  - Classification of document pages using various visual features that express "visual similarity" of layout structure
  - Decision tree classifiers and self-organizing maps
- Chen et al. (2006)
  - Image clustering as a basis for constructing visual words for representing documents
  - Bag-of-words representation and standard classification methods
  - Exploration of a new space of features, based purely on the clustering of subfigures for document classification

Related work (3/3)

Examples of studies that utilize both textual and visual features

- Liparas et al. (2014)
  - Classification of news articles using both textual and visual features
  - RF classifier
  - Late fusion strategy that exploits RF operational capabilities

- Augereau et al. (2014)
  - Classification of document images by combining textual and visual features
  - 1000 textual features extracted with the Bag of Words (BoW) method
  - 1000 visual features extracted with the Bag of Visual Words (BoVW) method

Feature extraction (1/3)

- Each article has two main p
  a) Images
  b) Textual content

- Extract N-gram features from the textual content
  - Globally
  - Relatively easy to compute
  - Effective for various classification tasks

- Select the biggest image of the article and extract low-level visual features
  - Assume that the biggest image is the representative one

Feature extraction (2/3)

N-gram textual features

- Delete all appearances of 421 stopwords for general texts in English (Fox, 1989)
- Create all possible continuous word N-gram (for N =1, 2, 3, 4)
- Count the frequency of each N-gram feature in the corpus
- Sort the word unigram, bigram, trigram and fourgram (each group alone) features in descending order

- To avoid unnecessarily large number of N-grams, we select 624 most frequent N-grams: 500 unigrams, 100 bigrams, 20 trigrams, and 4 fourgrams
- The motivation for these numbers is as follows: The larger the value of N is, the smaller the number of relatively frequent N-grams in the corpus is. According to the frequencies of the frequent N-grams, the reduction factor was determined to be 5.

Feature extraction (3/3)

Visual features

- Low-level visual features
- RGB-SIFT visual descriptor
  - Extension of SIFT
  - Captures more information and is able to better represent the image than SIFT
- Visual word assignment step after the feature extraction
  - K-means clustering is applied, in order to acquire the visual vocabulary
  - VLAD encoding
- 4000 visual features in total

Dataset

News documents written in English downloaded from a large number of news web-sites that belong to 4 categories:

- Health (187 documents)
- Lifestyle-Leisure (326 documents)
- Nature-Environment (447 documents)
- Politics (277 documents)

1237 documents in total

- Manual annotation of the web pages

Description of the ML methods

- **J48**
  - Improved variant of the C4.5 decision tree method
  - Attempts to account for noise and missing data
  - Deals with numeric attributes by determining where thresholds for decision splits should be placed
- **Random Forests (RF)**
  - Ensemble learning method for classification and regression
  - Construction of a set of decision trees
  - Two sources of randomness in the operational procedures of RF:
  - Each decision tree is grown on a different bootstrap sample drawn randomly from the training data
  - At each node split during the construction of a tree, a random subset of m variables is selected from the original set and the best split based on these m variables is used
- **Sequential Minimal Optimization (SMO)**
  - Algorithm for solving the optimization problem that occurs during the training of Support Vector Machines (SVM)
  - SMO divides this problem into a series of smallest possible sub-problems, which are then resolved analytically

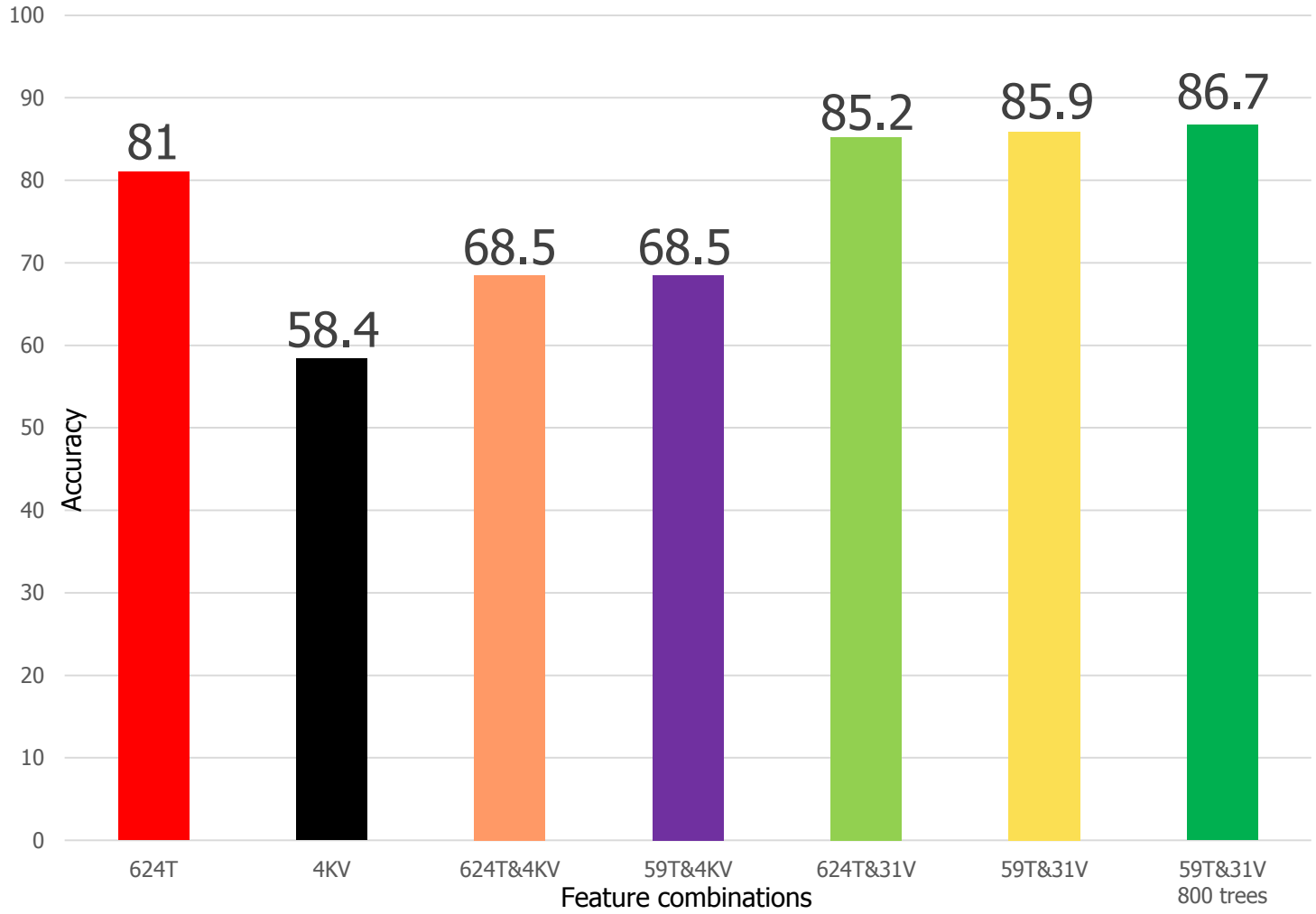
Experimental setup

- Experiments were conducted using the WEKA platform
- Default parameter values for the 3 ML methods
- 10-fold cross-validation
- Additional experiments that contain full parameter tuning for the two ML methods that gave the best initial results
- Application of a filter feature selection method
  - CfsSubsetEval (Correlation-based Feature Subset Selection)
  - Evaluates a subset of features by considering the individual predictive ability of each feature along

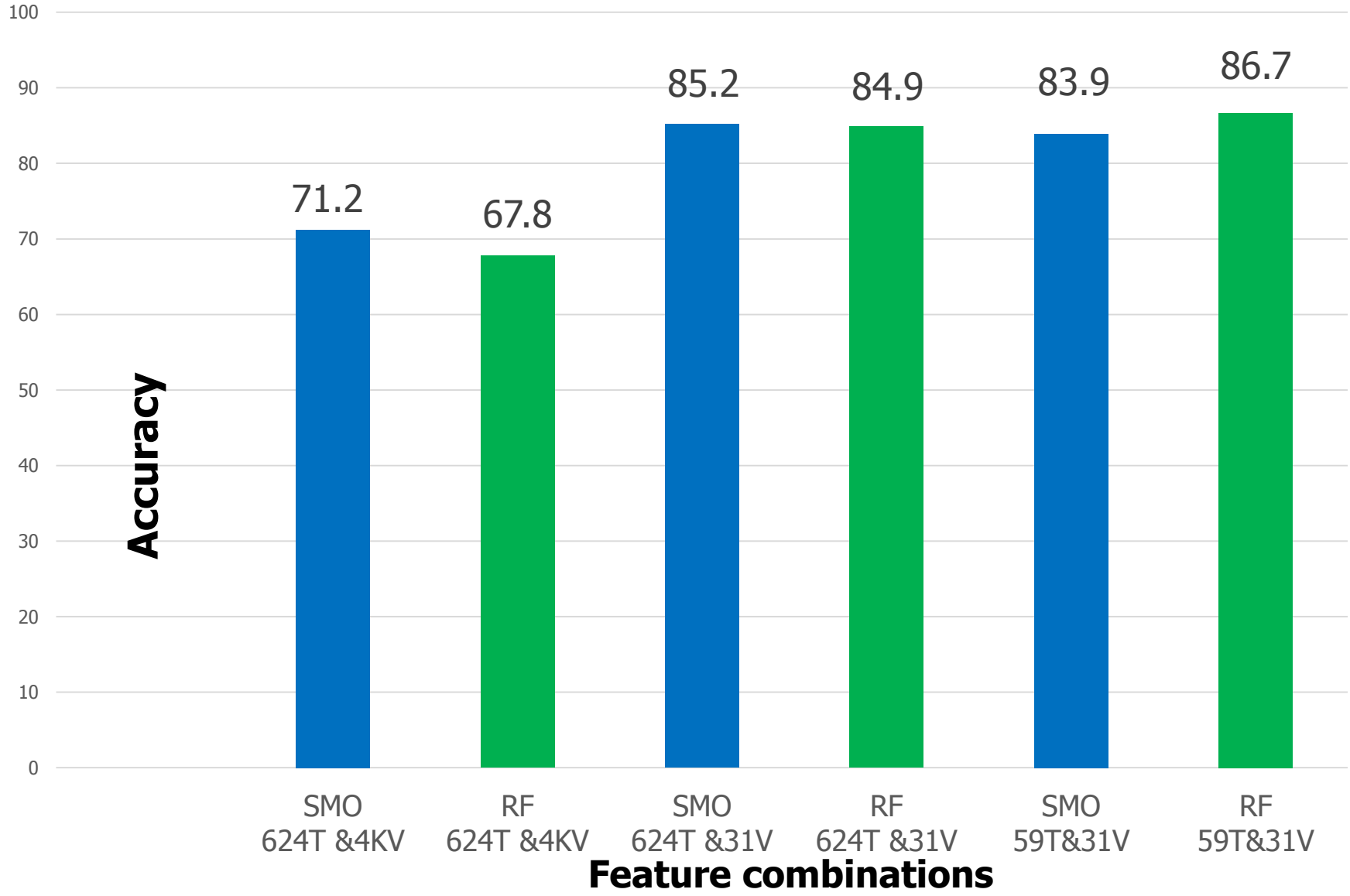
Accuracy results (%) for various combinations of feature sets using 3 ML methods with their default parameter values

Combinations of features	J48	RF	SMO
624 textual features	69.6	81.0	80.8
4000 visual features	55.1	58.4	57.3
59 textual features	72.9	82.3	79.8
31 visual features	54.2	59.4	52.1
624 textual & 4000 visual features	69.7	68.5	81.2
59 textual & 4000 visual features	69.5	68.5	80.8
624 textual & 31 visual features	68.4	85.2	83.9
59 textual & 31 visual features	72.7	**85.9**	82.9

Results Achieved by Random Forest (RF) using full parameter tuning



Results Achieved by RF & SMO using full parameter tuning



Summary and Conclusions (1/2)

- Comparative classification study of news documents
  - 3 popular ML methods (J48, RF and SMO)
  - Different combinations of key-phrases (word n-grams excluding stopwords) and visual features
- The use of N-gram textual features alone led to much better accuracy results (81.0%) than using only the visual features (58.4%)
- A possible explanation for this finding is that the textual features describe widespread information about the whole text, while the visual features describe information about only one representative image

Summary and Conclusions (2/2)

- Regarding RF:
  - Best combination of feature sets (59 textual and 31 visual features)
  - Best parameter values of RF: 800 trees and seed=3
  - Best accuracy result of 86.7%
  - Small accuracy improvement of 0.8% (from 85.9% to 86.7%) due to the parameter tuning

- Regarding SMO:
  - Best combination of feature sets (624 textual and 31 visual features)
  - Best parameter values: Normalized Polynomial Kernel,
    toleranceParameter = 0.003, c = 9
  - Best accuracy result of 85.2%
  - Small accuracy improvement of 1.3% (from 83.9% to 85.2%) due to the parameter tuning

Future work

- Define and implement additional types of textual features, e.g.
  - Function words
  - Morphological features (e.g.: nouns, verbs and adjectives)
  - Syntactic features (frequencies and distribution of PoS-tags)
- Apply additional ML methods based on textual and visual features
  - Larger number of documents
  - More categories
  - Other areas, applications and languages
- Selection of a representation for images based on **visual concepts**

Thank you very much