

# Language Identification by Using SIFT Features

Nikos Tatarakis , Ergina Kavallieratou

Dept. of Information and Communication System Engineering,  
University of the Aegean,  
Samos, Greece

**Abstract**—Two novel techniques for language identification of both, machine printed and handwritten document images, are presented. Language identification is the procedure where the language of a given document image is recognized and the appropriate language label is returned. In the proposed approaches, the main body size of the characters for each document image is determined, and accordingly, a sliding window is used, in order to extract the SIFT local features. Once a large number of features have been extracted from the training set, a visual vocabulary is created, by clustering the feature space. Data clustering is performed using K-means or Gaussian Mixture Models and the Expectation - Maximization algorithm. For each document image, a Bag of Visual Words or Fisher Vector representation is constructed, using the visual vocabulary and the extracted features of the document image. Finally, a multi class Support Vector Machine classification scheme is used, to score the system. Experiments are performed on well-known databases and comparative results with another established technique, are also given.

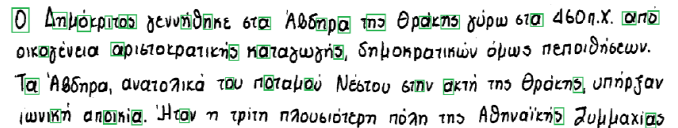
**Keywords**—Document image processing; language identification; SIFT features; bag of Visual Words; Fisher Vector

## I. INTRODUCTION

There is no doubt that nowadays all services tend to become even more digitalized. There is a growing trend for many multinational companies or organizations to digitalize their old documents or books and store them in digital databases. These documents could be either handwritten or machine printed and usually in various languages. Google Books [1] is an example. Language identification is a very important task, as it could help to index and/or sort a big digital document corpus in a convenient way.

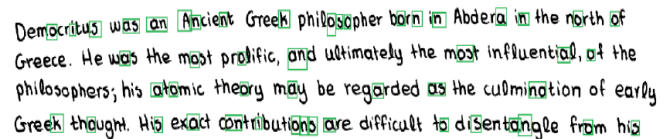
In the area of optical character recognition (OCR), most of the works assume that the language of the document is known beforehand. In this case, document language identification could be used as a valuable pre-processing task, in order to determine the correct language and eventually utilize the appropriate OCR engine for text extraction, translation and/or indexing.

The difficulty of language identification is highly dependent on the languages themselves. For example, a machine printed Chinese document is easily separated from a machine printed English document, since the local language structure, words and the shape of the letters, differ vastly in these two languages. Many papers have been published to address this issue.



Ο Δημόκριτος γεννήθηκε στα Αβδέρα της Θράκης γύρω στα 460 π.Χ. από οικογένεια αριστοκρατικής καταγωγής, δημοκρατικών όψεως πεποιδίσεων. Τα Αβδέρα, ανατολικά του ποταμού Νέστου στην ακτή της Θράκης, υπήρξαν ιωνική αποικία. Ήταν η τρίτη πλουσιότερη πόλη της Αθηναϊκής Συμμαχίας

Fig. 1. Handwritten Greek text



Democritus was an Ancient Greek philosopher born in Abdera in the north of Greece. He was the most prolific, and ultimately the most influential, of the philosophers; his atomic theory may be regarded as the culmination of early Greek thought. His exact contributions are difficult to disentangle from his

Fig. 2. Handwritten English text

In this paper, we focus on languages that are not very distant from each other and even share many common letters, like e.g. English and Greek. In Fig 1 and 2, common letters are noted, to show the very small intra-class variation.

To the best of our knowledge, most of the published work puts great emphasis on machine printed text that is also very linguistically diverse. The proposed works could be classified considering the methodology they are based on: a) Template Matching [2,3] b) Texture Analysis [4,5] and c) Shape Codebooks [6].

Hochberg et al. [2-3] presented a system for language identification matching cluster-based templates. In this work, they aim to discover repeating linguistic features like characters and word shapes in each script. To do so, they create fixed-size templates (textual symbols) from the training set, by clustering similar symbols together and representing each cluster by its centroid. They score the system by matching a subset of symbols to the templates. The reported results show high accuracy on a machine printed text corpus.

Texture analysis has also been proposed for the task of language identification. In [4], G. S. Peake et al. propose a segmentation-free approach. They used grey level co-occurrence matrices and Gabor filters in order to extract features. For classification, a K-NN classifier is used and they reported fair results on machine printed content. In [5], A. Busch et al. extracted wavelet co-occurrence features from small blocks of machine printed text and used a Gaussian Mixture Model (GMM) classifier to score the system. They reported good results, but on small machine printed textual regions and not on the whole document pages. More similar work on texture analysis can be found in [7, 8].

In [6], Zhu et al. propose a segmentation-free language identification system for both handwritten and machine printed documents for 8 languages. They extract local features that called Triple Adjacent Contours (TAS) from document images and form a shape codebook by clustering that feature space. They use that codebook to create an image descriptor for each document image. This descriptor is basically a statistical representation of the frequency that each TAS feature occurs on the image. They use a Support Vector Machine (SVM) for the classification process. They don't show how it performs on languages with low intra class-variation like Greek and English, but they report exceptional results on a wide language variety.

Apart from the aforementioned work, there are also a considerable number of other papers trying to address this problem, based mostly on line analysis of textual features. Projection profiles are being explored in [9] and upward concavities in [10]. These approaches usually require some preprocessing, like skew correction.

In this paper, two systems for the task of Language identification are presented. The first one is based on the Bag of Visual Features (BoVF) [11]. The second approach is based on representing the images using Fisher Vectors (FV) [12-13] created by a GMM visual vocabulary. For these two approaches, a Support Vector Machine classification scheme is used to score the system. Both methods share many similarities, however they perform differently.

The two systems are presented in the sections II and III, respectively, while the experimental results are analyzed in section IV. Finally our conclusions are drawn in section V.

## II. LANGUAGE IDENTIFICATION USING BAG OF VISUAL FEATURES

The Bag of Visual Features (BoVF) is one of the suggested methods for generic Image categorization based on image content. This particular method is widely applied in scene and object categorization as well as image retrieval [14-15]. The Bag of Visual Features has been inspired from the Bag of Words of the area of Natural Language Processing that is used for text classification. Generally speaking, the mentioned systems are trying to extract relevant features from the images in order to create a histogram representation of every image (Fig.3) that will be used as input to a classifier.

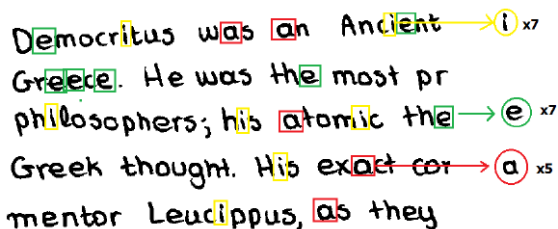


Fig. 3. To represent an image as a Bag of Visual Features, each feature is mapped to the nearest image feature/word cluster. All the information of each document image is finally included in e.g. 300 features, out of the e.g. 500-1000 clusters of Visual Features

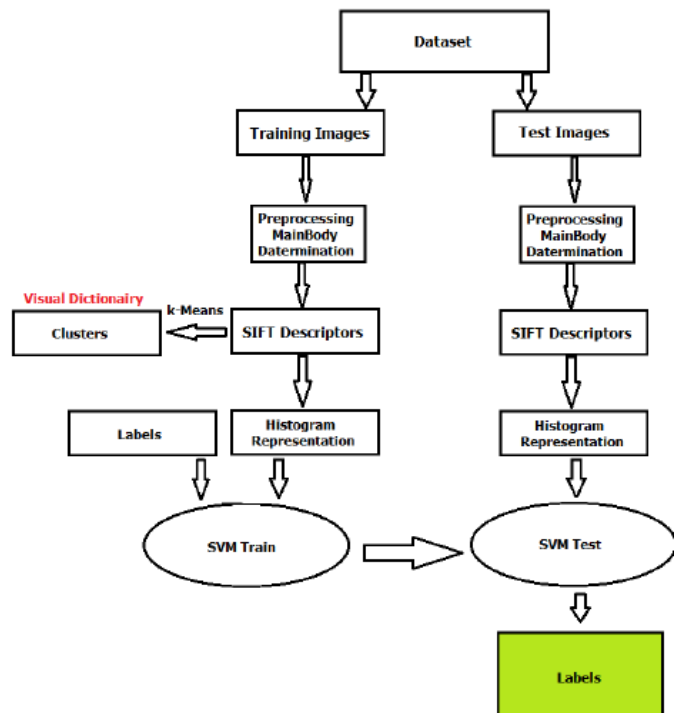


Fig. 4. The Bag of Visual Features approach

The BoVF system that we implement consists of the following steps:

- 1) A Preprocessing step for every document image that determines the Main Body size of the characters.
- 2) Feature Extraction for every document image.
- 3) Creation of a visual vocabulary using k-means.
- 4) Histogram creation for every document image, which represents the frequency of the visual features in the image.
- 5) Training of an SVM classification scheme that will be fed with the histograms. The BoVF is a method that only the feature appearance frequency is used from the image without taking into consideration the natural position of every feature on the image. However, in our application, the position of the objects is not important. In Fig.4, the proposed Bag-of-Visual-Feature approach is shown.

The first step consists of the application of the Main Body detection algorithm, described in [16], to the document image. This will return a number  $n$  which indicates how tall is the main body of the majority of the characters in the image. During the feature extraction procedure, this information is used in order to create a  $nxn$  sliding window to dense sample the image for SIFT features using the algorithm in [17].

- 1) The SIFT Descriptor is a spatial histogram with  $4x4$  bins. By setting the size of sliding window to  $nxn$ , which is given by the main body detection algorithm, actually the SIFT descriptor is computed in a  $4nx4n$  pixel area (Fig.5).

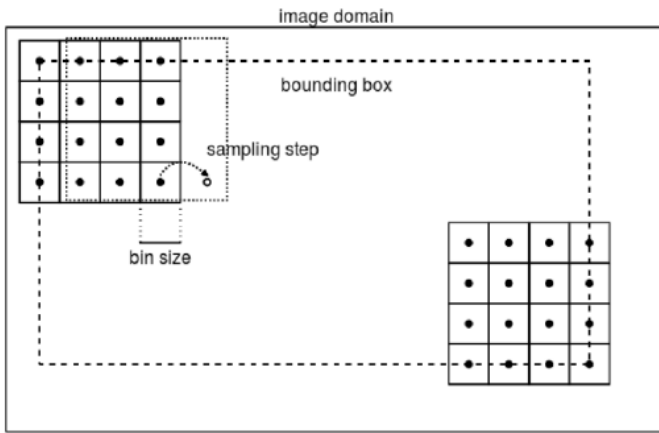


Fig. 5. Dense SIFT descriptor geometry [18]

This step proved to be particularly important as the document images within a class (same language) have quite similar main body estimations. This scheme creates a more personalized and more sophisticated local descriptor within a class, which creates more unique patterns inside the data and maximizes the inter-class variation, while at the same time the intra-class variation is minimized.

The SIFT descriptor is used, since it is sufficiently robust against noise, as well as scale and rotation invariant, which is very important, especially when we have to deal with badly scanned document images. Finally, the SIFT descriptor is 128-dimensional, which maintains as much information as possible and a more distinctive image representation is achieved.

The visual vocabulary can be thought as a big dictionary that contains visual words (features) from the training set. After a feature is extracted from an image, the visual vocabulary is used in order to map this image feature to the closest cluster. An extreme approach would be to compare every image feature to every feature from the training space. However, this is quite unlikely to happen as the feature space from the training set is extremely large and such a task would be computationally expensive. In order to shrink the feature space and create a comprehensive codebook, k-means is used. As usually, it is hard to define how many clusters are enough for the task. Tests with several k values have to be performed to decide, which the correct value for our data is (section 4).

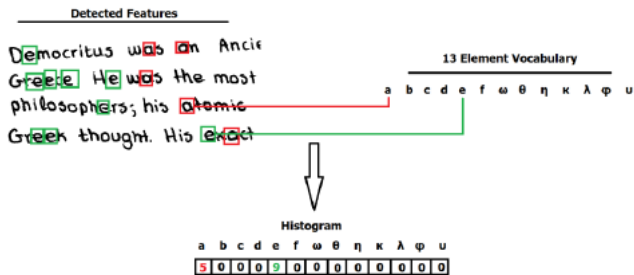


Fig. 6. Histogram creation

Once the visual vocabulary is created, the image representation follows. For every image, the features are extracted. For every feature, the closest entry in the visual vocabulary is detected. The Euclidean distance is used to

define how close the visual feature to each cluster is. The size of the resulting histogram for each image equals to the size of the dictionary (Fig.6). However, the histogram is usually very sparse, since most of the words of the visual vocabulary are too distant to match with an image feature, especially if they belong to a different language. As a result, some bins of the histogram are overpopulated, while all the others have small or null values. This high variance could cause trouble to the classifier so the histogram is normalized to 1.

The Classification is the last task of the proposed system. An one- against -all SVM classification is used. The idea behind one-against-all approach is, to train Support Vector Machine models each one separating one class from the rest. Every time, the classifier with the largest decision value is chosen. During training, the classifier is fed with the training histograms and their labels (1/-1). During the testing process, the test histograms are provided and the predicted label is received.

### III. LANGUAGE IDENTIFICATION USING FISHER VECTOR

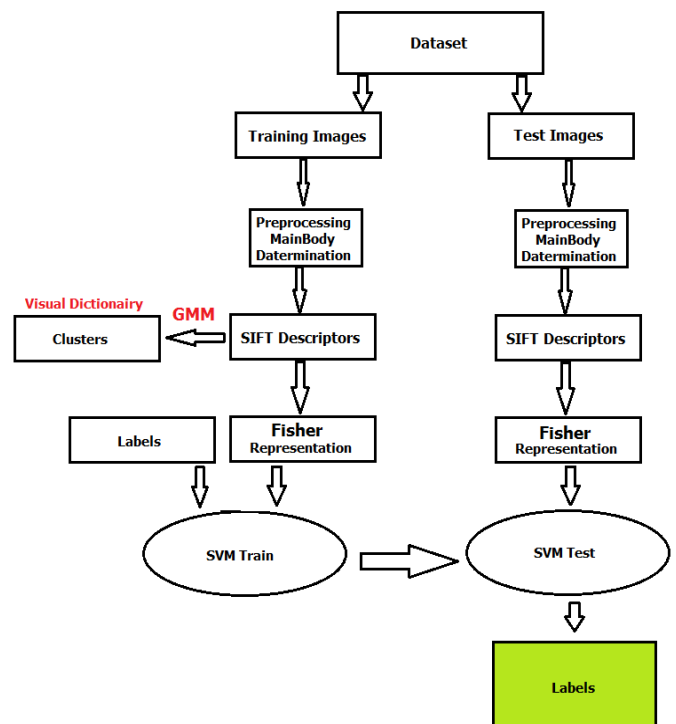


Fig. 7. The Fisher Vector approach

The Fisher Vector (FV) system is basically an extension to the Bag of Visual Features approach. However, it differs a lot, when it comes to create the visual vocabulary and to represent the images. Here, the visual vocabulary is created by training clusters using the Gaussian Mixture Models (GMM). Following the clustering, the images are represented by Fisher Vectors. In this section, the procedure that builds a Fisher vector, using the clustering given by a GMM, is presented. The other tasks have been analyzed in the previous section.

Thus, Fisher Vector pipeline consists of the following steps:

- A Preprocessing step for every document image that help us to determine the Main Body size of the characters.
- Feature Extraction for every document image.
- Creation of a visual vocabulary using Gaussian Mixture Models.
- Fisher Vector representation is created for every document image.
- Training of an SVM classification scheme that will be fed with the Fisher vectors.

In Fig.7 the proposed approach is shown.

In [13], a new algorithm for image representation using the FV is introduced, given local image descriptors and a GMM probabilistic vocabulary. This algorithm as it is used in our approach is described in Fig.8. The dimensionality of this image descriptor is  $K(2D+1)$ , where K is the number of GMM clusters and D the dimensions of the local image descriptor. Thus, the D is 128, due to the SIFT descriptor, but the K also needs to be determined through experiments (section 4).

**Input:**

- Local image descriptors  $X = \{x_t \in R^D, t = 1, \dots, T\}$ ,
- Gaussian Mixture model parameters  $\lambda = \{w_k, \mu_k, \sigma_k, k = 1, \dots, K\}$

**Output:**

- Normalized Fisher Vector representation  $G_\lambda^X \in R^{K(2D+1)}$

**1. Compute Statistics**

- For  $k = 1, \dots, K$  initialize accumulators
  - $S_k^0 = 0, S_k^1 = 0, S_k^2 = 0$
- For  $t = 1, \dots, T$ 
  - Compute  $\gamma_t(k)$  using (3.4.1.1.1 formula, 6)
  - For  $k = 1, \dots, T$ 
    - $S_k^0 = S_k^0 + \gamma_t(k)$ ,
    - $S_k^1 = S_k^1 + \gamma_t(k)x_t$ ,
    - $S_k^2 = S_k^2 + \gamma_t(k)x_t^2$

**2. Compute Fisher Vector Signature**

- For  $k = 1, \dots, K$ :
 
$$G_{\alpha_k}^X = (S_k^0 - Tw_k) / \sqrt{w_k}$$

$$G_{\beta_k}^X = (S_k^1 - \mu_k S_k^0) / \sqrt{w_k \sigma_k}$$

$$G_{\sigma_k}^X = (S_k^2 - 2\mu_k S_k^1 + (\mu_k^2 - \sigma_k^2) / \sqrt{2w_k \sigma_k^2})$$
- Concatenate all Fisher vector components into one vector
 
$$G_\lambda^X = (G_{\alpha_1}^X, \dots, G_{\alpha_K}^X, G_{\beta_1}^X, \dots, G_{\beta_K}^X, G_{\sigma_1}^X, \dots, G_{\sigma_K}^X)$$

**3. Apply normalizations**

- For  $i = 1, \dots, K(2D+1)$  apply power normalization
 
$$[G_\lambda^X]_i \leftarrow \text{sign}([G_\lambda^X]_i) \sqrt{|[G_\lambda^X]_i|}$$
- Apply  $l_2$  - normalization
 
$$G_\lambda^X = G_\lambda^X / \sqrt{G_\lambda^X G_\lambda^X}$$

Fig. 8. The algorithm for the Fisher Vector system

#### IV. EXPERIMENTAL RESULTS

In order to evaluate these systems, three different datasets are used. The first is a labmade dataset that consists of machine printed text in Greek, English and Arabic document images that have been taken randomly from papers and e-books all over the web. It contains 120 test images, 40 from each language, and 180 training images, 60 from each language. This is a high resolution dataset. The second is the dataset of ICDAR2013 Handwriting Segmentation Contest [19]. This one also consists of high resolution images of handwritten text in English, Greek and Indian. It contains 150 test images, 50 from each language, and 200 training images, 75 from each language. The third dataset was intended to be as close as possible to the one described in [6], in order to give comparative results. In order To evaluate the performance of our multiclass SVM classifier and to score the system, confusion matrixes and the accuracy rate were used.

##### A. The Bag of Visual Feature System

First, it examined how the results are affected by the window size. The goal is to determine the window size in relation to the main body size, as a natural normalization of the writing style. Once the optimal size is determined, then, additional experiments will be performed in order to evaluate the size of the vocabulary.

Roughly, around 600-700 SIFT features are extracted from each image. This step of the algorithm is particularly time consuming, since the images have been taken in high resolution. However, trying to keep preprocessing as low as possible, image resizing techniques that would probably cause valuable information loss were not applied.

In this experiment, the size of the codebook was kept relatively small at the fixed value of  $k=50$ , while performing experimenting with the following rectangular window sizes: main body x 1, main body x 2 and round(main body x 0.5). In tables 1-3, the results are shown on Labmade dataset, while in Fig.9 the accuracy graph for those values is presented.

TABLE I. DATA SET= LABMADE, K=50, WINDOW=MAIN BODY X 1, SVM C PARAMETER = 0.0001

		Predicted Class		
		Greek	English	Arabic
Labmade Data Set				
Actual Class	Greek	0.525	0.35	0.125
	English	0.3	0.55	0.15
	Arabic	0.175	0.125	0.70
Mean Diagonal Accuracy : 59.16%				

TABLE II. DATA SET= LABMADE, K=50, WINDOW = MAIN BODY X 2 , SVM C PARAMETER = 0.0001

Labmade Data Set		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.475	0.325	0.20
	English	0.35	0.525	0.125
	Arabic	0.2	0.125	0.675
<b>Mean Diagonal Accuracy : 55.83%</b>				

TABLE III. DATA SET= LABMADE, K=50, WINDOW =ROUND(MAIN BODYX0.5), SVM C PARAMETER = 0.0001

Labmade Data Set		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.45	0.35	0.20
	English	0.375	0.475	0.15
	Arabic	0.175	0.15	0.675
<b>Mean Diagonal Accuracy : 53.33%</b>				

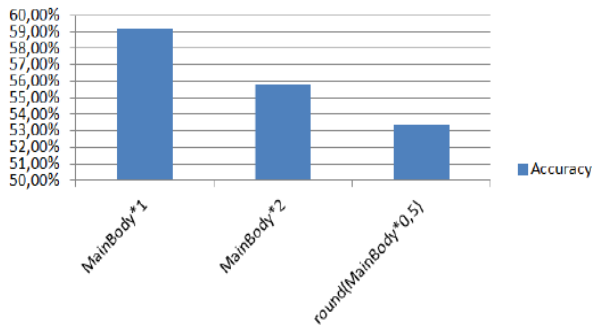


Fig. 9. Collective Accuracy Graph

It is clear that the highest accuracy was obtained by using the rectangular window with size main bodyx1. The results look quite low, since the visual vocabulary contained 50 visual words, only. In the following experiments the sliding window will constantly set at this size, while the vocabulary size will be increased from 100 to 2000 features. The results of this experiment are presented in tables 4-8, while the accuracy graph is shown in Fig.10.

TABLE IV. DATA SET= LABMADE, K=100, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

Labmade Data Set		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.65	0.225	0.125
	English	0.175	0.675	0.15
	Arabic	0.2	0.075	0.725
<b>Mean Diagonal Accuracy : 68.33%</b>				

TABLE V. DATA SET= LABMADE, K=200, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

Homemade Data Set		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.70	0.20	0.10
	English	0.175	0.75	0.075
	Arabic	0.10	0.075	0.825
<b>Mean Diagonal Accuracy : 75.83%</b>				

TABLE VI. DATA SET= LABMADE, K=500, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

Homemade Data Set		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.8	0.15	0.05
	English	0.125	0.825	0.05
	Arabic	0.075	0.025	0.9
<b>Mean Diagonal Accuracy : 84.16%</b>				

TABLE VII. DATA SET= LABMADE, K=1000, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

Labmade Data Set		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.85	0.125	0.025
	English	0.1	0.875	0.025
	Arabic	0.025	0.025	0.95
<b>Mean Diagonal Accuracy : 89.16%</b>				

TABLE VIII. DATA SET= LABMADE, K=2000, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

Labmade Data Set		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.875	0.075	0.05
	English	0.075	0.90	0.025
	Arabic	0.025	0.025	0.95
<b>Mean Diagonal Accuracy : 90.83%</b>				

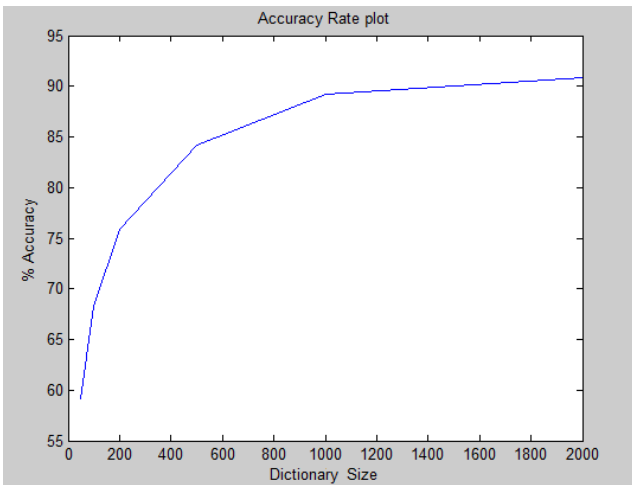


Fig. 10. Accuracy vs. Dictionary size

Apparently, the dictionary size is a very important parameter because it highly affects the overall performance. For the Labmade dataset, the highest score was achieved by using a relatively large dictionary of 2000 centers. It is worth noting that after 1000 centers the accuracy is getting marginally better. Therefore, anything over 2000 is expected to increase performance cost rather than the accuracy.

TABLE IX. DATA SET= ICDAR 2013, K=2000, WINDOW = MAIN BODY, SVM C PARAMETER = 0.0001

ICDAR 2013		Predicted Class		
		Greek	English	Indian
Actual Class	Greek	0.86	0.12	0.02
	English	0.14	0.84	0.02
	Indian	0.04	0.02	0.94
<b>Mean Diagonal Accuracy : 88%</b>				

Similarly, on the ICDAR 2013 dataset the best results obtained for window size = *main body x 1* and  $k=2000$  (table 9).

Regarding ICDAR2013 dataset, the best performance for the Bag Of Visual Features model is a bit lower than the best result of the labmade dataset, since this dataset is much harder as it contains only handwritten documents images.

#### B. The Fisher Vector System

For this approach, the optimal dictionary size among our experiments on the ICDAR dataset proved to be 256 elements (tables 10-12, Fig.11), since for more elements the whole system gets very slow and eventually it runs out of memory. Similarly, in [20] they also suggest a codebook of 256 clusters. Again, the window of size=*main body x 1* was used. Finally, in table 13, the results for the Labmade dataset are given.

TABLE X. DATA SET= ICDAR2013, K=64, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

ICDAR 2013		Predicted Class		
		Greek	English	Indian
Actual Class	Greek	86	10	4
	English	12	86	2
	Indian	2	0	98
<b>Mean Diagonal Accuracy : 90%</b>				

TABLE XI. DATA SET= ICDAR2013, K=128, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

ICDAR 2013		Predicted Class		
		Greek	English	Indian
Actual Class	Greek	88	12	0
	English	8	92	0
	Indian	2	0	98
<b>Mean Diagonal Accuracy : 92.66%</b>				



TABLE XII. DATA SET= ICDAR2013, K=256, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

		Predicted Class		
		Greek	English	Indian
Actual Class	Greek	0.92	0.08	0
	English	0.06	0.94	0
	Indian	0	0	1
<b>Mean Diagonal Accuracy : 95,33%</b>				

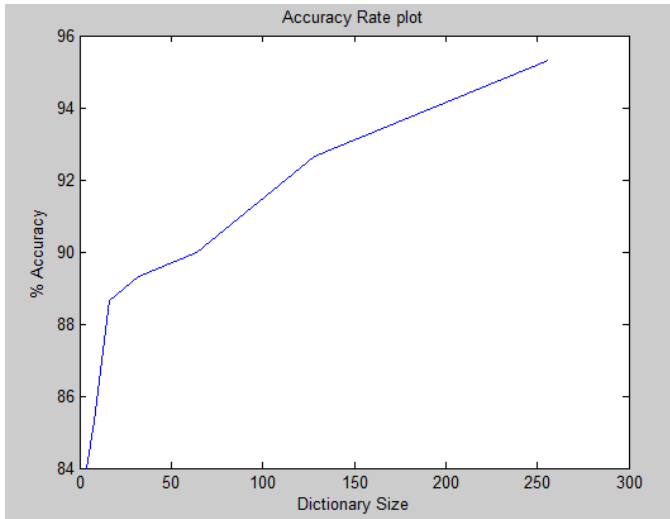


Fig. 11. Accuracy vs. Dictionary size

TABLE XIII. DATA SET= LABMADE, K=256, WINDOW = MAIN BODY X 1, SVM C PARAMETER = 0.0001

		Predicted Class		
		Greek	English	Arabic
Actual Class	Greek	0.925	0.05	0.025
	English	0.025	0.95	0.025
	Arabic	0.025	0	0.975
<b>Mean Diagonal Accuracy : 95%</b>				

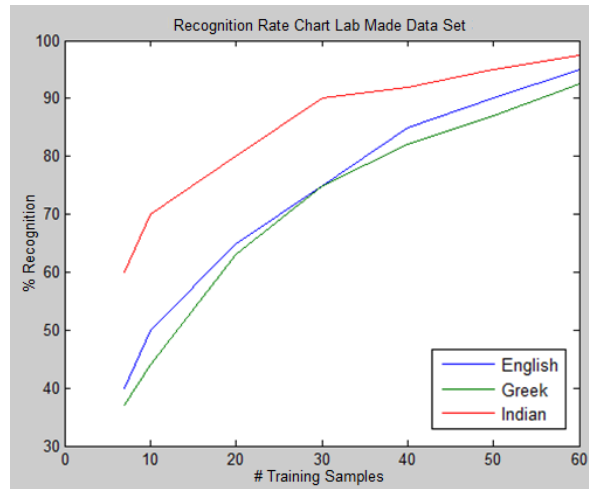


Fig. 12. Recognition Rate for the Labmade dataset

These results are quite surprising since they give higher scores in ICDAR 2013 dataset, which is objectively a harder handwritten dataset compared to the Labmade. However, by a closer look on the data in the labmade results, the scores regarding the Greek and English documents are slightly higher than their handwritten equivalent ones (Fig.12-13).

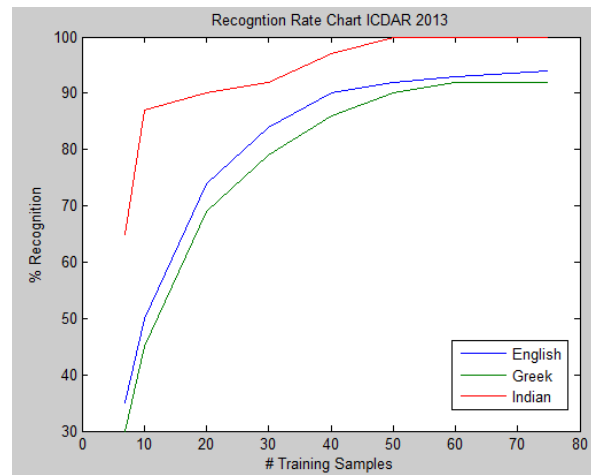


Fig. 13. Recognition Rate for the ICDAR 2013

This proves that languages with very distinctive features like Arabic and Indian are much easier to separate. Even with a small amount of training samples it is enough to obtain acceptable recognition rates.

It was mentioned that the dimensionality of a Fisher Vector representation is given by the formula  $K(2D+1)$ , where  $K$  is the number of GMM clusters and  $D$  the local image descriptor dimensions. Suppose that we have formed a GMM dictionary of  $K=256$  centers like above. As the SIFT descriptor is 128-

dimensional, every image is represented by a vector of  $256 \times (2 \times 128 + 1)$ , or 65792 elements. A good idea to shrink the Fisher vector is the application of PCA to the 128-D SIFT descriptors [20]. This would lead to a more computationally efficient classification scheme as the SVM would have to deal with smaller Fisher Vectors. In [21], they propose to apply PCA on SIFT descriptors to reduce their dimension from 128 to 64. This would cut down the final Fisher Vector to 33024 elements instead of 65792, which makes the system much more efficient and less memory starving. In [20] they report some increase in their overall classification performance after applying PCA, about 4%. In our system, regarding the effect of PCA, apart of the faster computation and less memory demands, the overall performance was slightly dropped (Fig.14).

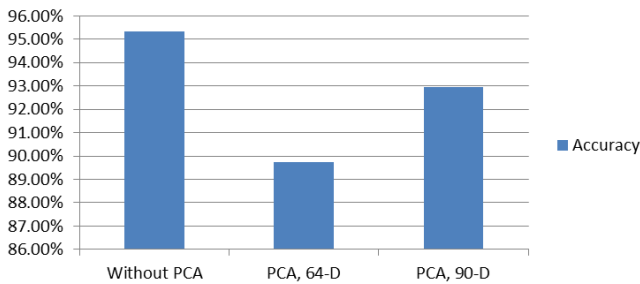


Fig. 14. Considering PCA for the proposed system

In any case, the reported results were already pretty high, so it was quite expected that PCA wouldn't help drastically. In fact, it didn't help at all. This also could mean that the extracted features are of high quality and very discriminative and contain very small amounts of redundant data, if any at all.

C. Comparative results

In this section, comparative results to the system presented in [6] are given.

Maryland + IAM 3.0		Predicted Class (Percentages)							
		Thai	Cyrillic	Chinese	Hindi	Korean	Japanese	Arabic	English
Actual Class	Thai	100	0	0	0	0	0	0	0
	Cyrillic	0	97.5	0	0	0	0	2.5	0
	Chinese	0	0	97.5	0	2.5	0	0	0
	Hindi	0	0	2.5	97.5	0	0	0	0
	Korean	0	0	0	0	100	0	0	0
	Japanese	0	0	6.65	0	0	66.7	26.65	0
	Arabic	0	0	0	0	8.70	0	91.30	0
	English	0	1	0	0	0	0	0.5	98.5
Mean Diagonal Accuracy : 93.62%									

Fig. 15. Confusion matrix for the BoFV system

Maryland + IAM 3.0		Predicted Class (Percentages)							
		Thai	Cyrillic	Chinese	Hindi	Korean	Japanese	Arabic	English
Actual Class	Thai	100	0	0	0	0	0	0	0
	Cyrillic	0	100	0	0	0	0	0	0
	Chinese	0	0	97.5	0	2.5	0	0	0
	Hindi	0	0	0	100	0	0	0	0
	Korean	0	0	0	0	100	0	0	0
	Japanese	0	0	0	0	0	80	20	0
	Arabic	0	0	0	0	0	0	100	0
	English	0	0	0	0	0	0	0	100
Mean Diagonal Accuracy : 97.18%									

Fig. 16. Confusion matrix for the FV system

Maryland + IAM 3.0		Predicted Class (Percentages)							
		Thai	Cyrillic	Chinese	Hindi	Korean	Japanese	Arabic	English
Actual Class	Thai	96.3	0	0.3	0.9	0.3	0.6	0	1.6
	Cyrillic	0.4	97.1	0	0	0	0	0.5	2
	Chinese	0.2	0.7	85	1	1	6.7	1.4	4
	Hindi	0	0	0.2	98.8	0	0.8	0.2	0
	Korean	0.1	0.5	0.8	1.9	96	0.5	0	0.1
	Japanese	0	0	1.3	0.2	1.3	96.2	0	1
	Arabic	0	0	0.3	0	0	0	99.7	0
	English	0.6	0.6	0	0.2	1.1	0	1.6	95.9
Mean Diagonal Accuracy : 95.6%									

Fig. 17. Confusion matrix for the system in [6]

To evaluate objectively the proposed systems, they are compared to the current state of the art system presented by G. Zhu et al. in [6] by using the same databases; The IAM handwriting DB3.0 database [22] as well as the University of Maryland multilingual database [23]. The dataset contains over 1000 document images from both databases in 8 languages (Thai, Cyrillic, Chinese, Hindi, Korean, Japanese, Arabic and English). The comparative results are given (Fig.15-17) in the form of confusion tables.

V. CONCLUSIONS

In this paper, two systems for the task of Language Identification based on document Images have been proposed. To evaluate the performance of the proposed systems, three datasets were used. From the experimental results, it is concluded that both of these systems have high potential, although the Fisher Vector approach proved to be much better and more accurate than the Bag of Visual Features. It is also safe to claim that the Fisher Vector method is able to outperform the current state of the art approach for the task of Language Identification as it is shown in section 4.3.



The main advantages of the Fisher Vector (FV) image representation over Bag of Visual Features (BoVF) representation, is, that the former, using a much smaller dictionary, that contains only 256 clusters, it significantly outperforms the BoVF system which needs at least a 1000-words dictionary to perform well, as it was proved in our experiments. Therefore, the FV system has a lower computational cost as it requires a much smaller dictionary to obtain acceptable recognition accuracy. The only issue with Fisher vectors is that they are quite dense compared to the sparse BoVF histograms, which makes them unappealing, in case we have to deal with a lot of data, as the storage and I/O requirements will increase dramatically.

Another very important fact is that even with a small number of training samples, it can perform exceptionally well, especially if the document language includes very distinctive features.

Finally, the choice to include a preprocessing task like the main body estimation is of high importance. This task helped to extract very distinctive, personalized and relevant features from every image and increased the overall accuracy of our system.

In the future, we are going to experiment with more techniques inspired from Natural Language Processing for the document Image Processing tasks.

#### REFERENCES

- [1] L. Vincent, "Google book search: Document understanding on a massive scale", In Proc. Int. Conf. Document Analysis and Recognition, pages 819–823, 2007.
- [2] J. Hochberg, L. Kerns, P. Kelly, T. Thomas, "Automatic script identification from images using cluster-based templates", Proceedings of Third International Conference on Document Analysis and Recognition, Vol.1, pp. 378, 1995.
- [3] J. Hochberg, P. Kelly, T. Thomas, L. Kerns, "Automatic Script Identification From Document Images Using Cluster-Based Templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 176-181, 1997.
- [4] G. S. Peake, T.N. Tan, "Script and Language Identification from Document Images", Proceedings of the Workshop on Document Image Analysis, pp.10-17, 1997.
- [5] A. Busch, W. W. Boles, and S. Sridharan, "Texture for script identification" IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(11):1720–1732, 2005.
- [6] G. Zhu, X. Yu, Y. Li, and D. Doermann, "Unconstrained Language Identification Using A Shape Codebook", ICFHR2008, p.13–18, 2008.
- [7] T. Tan, "Rotation Invariant Texture Features And Their Use In Automatic Script Identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(7):751–756, 1998.
- [8] V. Singhal, N. Navin, and D. Ghosh, "Script-based classification of hand-written text document in a multilingual environment," in Proc. of Int. Workshop on Research Issues in Data Eng., 2003, pp. 47–54. -
- [9] S. L. Wood, Xiaozhong Yao, K. Krishnamurthi, L. Dang, Language identification for printed text independent of segmentation, Proceedings of the International Conference on Image Processing, vol.3, pp.3428-3431, 1995. -
- [10] A. L. Spitz, "Determination of the Script and Language Content of Document Images", IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 235-245, 1997.
- [11] G.Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, "Visual categorization with bags of keypoints", ECCV'04 workshop on Statistical Learning in Computer Vision, pp. 59-74, 2004.
- [12] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In CVPR, 2007. -
- [13] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice", IJCV, 105(3):222–245, June 2013.
- [14] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In International Conference on Computer Vision, 2005. =
- [15] J. Liu. Image retrieval based on bag-of-words model. In arXiv preprint arXiv:1304.5168, 2013. -
- [16] P. Diamantatos, V. Verras, E. Kavallieratou. Detecting Main Body Size in Document Images. ICDAR, 2013.
- [17] D.G.Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60:91–110, 2004.
- [18] A. Vedaldi and B. Fulkerson. VLFeat library. <http://www.vlfeat.org/>, 2008.
- [19] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal, A. Alaei, "ICDAR 2013 handwriting segmentation contest", 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1402-1406, 2013.
- [20] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In BMVC, 2011.
- [21] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In Proc. CVPR, 2004. L. Vincent, "Google book search: Document understanding on a massive scale", In Proc. Int. Conf. Document Analysis and Recognition, pages 819–823, 2007.
- [22] J. Hochberg, L. Kerns, P. Kelly, T. Thomas, "Automatic script identification from images using cluster-based templates", Proceedings of Third International Conference on Document Analysis and Recognition, Vol.1, pp. 378, 1995.
- [23] J. Hochberg, P. Kelly, T. Thomas, L. Kerns, "Automatic Script Identification From Document Images Using Cluster-Based Templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 176-181, 1997.
- [24] G. S. Peake, T.N. Tan, "Script and Language Identification from Document Images", Proceedings of the Workshop on Document Image Analysis, pp.10-17, 1997.
- [25] A. Busch, W. W. Boles, and S. Sridharan, "Texture for script identification" IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(11):1720–1732, 2005.
- [26] G. Zhu, X. Yu, Y. Li, and D. Doermann, "Unconstrained Language Identification Using A Shape Codebook", ICFHR2008, p.13–18, 2008.
- [27] T. Tan, "Rotation Invariant Texture Features And Their Use In Automatic Script Identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(7):751–756, 1998.
- [28] V. Singhal, N. Navin, and D. Ghosh, "Script-based classification of hand-written text document in a multilingual environment," in Proc. of Int. Workshop on Research Issues in Data Eng., 2003, pp. 47–54. -
- [29] S. L. Wood, Xiaozhong Yao, K. Krishnamurthi, L. Dang, Language identification for printed text independent of segmentation, Proceedings of the International Conference on Image Processing, vol.3, pp.3428-3431, 1995. -
- [30] A. L. Spitz, "Determination of the Script and Language Content of Document Images", IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 235-245, 1997.
- [31] G.Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, "Visual categorization with bags of keypoints", ECCV'04 workshop on Statistical Learning in Computer Vision, pp. 59-74, 2004.
- [32] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In CVPR, 2007. -
- [33] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice", IJCV, 105(3):222–245, June 2013.
- [34] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In International Conference on Computer Vision, 2005. =

- [35] J. Liu. Image retrieval based on bag-of-words model. InarXiv preprint arXiv:1304.5168, 2013. -
- [36] P. Diamantatos, V. Verras, E. Kavallieratou. Detecting Main Body Size in Document Images. ICDAR, 2013.
- [37] D.G.Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60:91–110, 2004.
- [38] A. Vedaldi and B. Fulkerson. VLFeat library. <http://www.vlfeat.org/>, 2008.
- [39] N.Stamatopoulos, B.Gatos, G.Louloudis, U.Pal, A.Alaei, "ICDAR 2013 handwriting segmentation contest", 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1402-1406, 2013.
- [40] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In BMVC, 2011.
- [41] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In Proc. CVPR, 2004.