DETECTION OF PORNOGRAPHIC IMAGES USING BAG-OF-VISUAL-WORDS AND ARCX4 OF RANDOM MULTINOMIAL NAIVE BAYES

Thanh-Nghi Do[1,2]

[1]*Institut Telecom; Telecom BretagneUMR CNRS 3192 Lab-STICC*

[2]*Universite europeenne de Bretagne, France Can Tho University, Vietnam*

Received September 30, 2011

ABSTRACT

The paper presents a novel approach to detect pornographic images. At the pre-processing step, we propose to use the Scale-invariant feature transform method (SIFT) which is locally based on the appearance of the object at particular interest points, invariant to image scale, rotation and also robust to changes in illumination, noise, occlusion. And then, the representation of the image that we use for classification is the bag-of-visual-words (BoVW), which is constructed from the local descriptors and the counting of the occurrence of visual words in a histogram like fashion. The pre-processing step brings out datasets with a very large number of dimensions. And then, we propose a new algorithm called Arcx4 of random multinomial naive Bayes (Arcx4-rMNB) that is suited for classifying very-high-dimensional datasets. We do setup experiment with two real datasets to evaluate performances. Our approach has achieved an accuracy of 91.75% for a small dataset and 87.93% for other large one.

1. INTRODUCTION

Since Internet grows quickly, people benefit more and more from the sharing of information. Thus, pornography increases rapidly and is one of the highly distributed information over web. It may be harmful to children. Therefore protecting our kids from exposure to pornographic images is a very pressing issue in the real world. Due to this aims, the researchers study the approaches for the recognition of pornographic images, i.e. learning image content to detect pornographic images.

There two major approaches to detect pornographic images. The first one is based on the detection of skin color pixels, skin texture and color histograms [6 - 8, 19, 24] and faces [10], body shape [22]. These systems use neural networks, support vector machines [21] or random forests [3] for learning to classify pornographic images. However, the first approach does not achieve a high accuracy. Recently, the second one [5, 12] is based on the Scale-invariant feature transform method (SIFT [13]) and the bag-of-visual-words (BoVW) models (initially proposed by [1] for texture classification). An image is represented by a set of visual words, which are obtained by clustering local descriptors. The pre-processing step brings out datasets with a very large number of dimensions (e.g. 2000 dimensions or visual words). And then, support vector machines [21] are usually suited for classifying this kind of data. This approach usually outperforms the first one.

In this paper, we also propose to use the SIFT and the BoVW model for representing images, furthermore we investigate a new machine learning algorithm called Arcx4 of random multinomial naive Bayes (Arcx4-rMNB) for classifying pornographic images. Our Arcx4-rMNB algorithm uses Arcx4 approach [2] to constructs sequentially k random multinomial naive Bayes so that each weak classifier (i.e. random multinomial naive Bayes) concentrates mostly on the errors produced by the previous ones. Furthermore, we propose to use random subset features to estimate class probabilities of multinomial naive Bayes. Thus this idea increases noise robustness of multinomial naive Bayes. Therefore, the ArcX4-rMNB can deal with very-high-dimensional dataset (many input features with each one containing only a small amount of information) like the image representation using BoVW models. The numerical test results on two real datasets of images showed that our proposal has achieved an accuracy of 91.75% for a small dataset and 87.93% for other large one. Our Arcx4-rMNB algorithm outperforms the state-of-the-art algorithms including decision tree C4.5 [17], random forest of CART (RF-CART [3]), AdaBoost of C4.5 [9]. In comparison with SVM [21], our Arcx4-rMNB outperforms SVM in terms of correctness rate of *porno* but SVM gives best results in terms of correctness rate of *non — porno.*

The paper is organized as follows. Section 2 presents the image representation using the SIFT and the BoVW model. Section 3 briefly introduces multinomial naive Bayes and our Arcx4-rMNB algorithm for classification of very-high-dimensional datasets. The experimental results are presented in Section 4. We then conclude in Section 5.

2. SIFT AND BAG-OF-VISUAL-WORDS MODEL

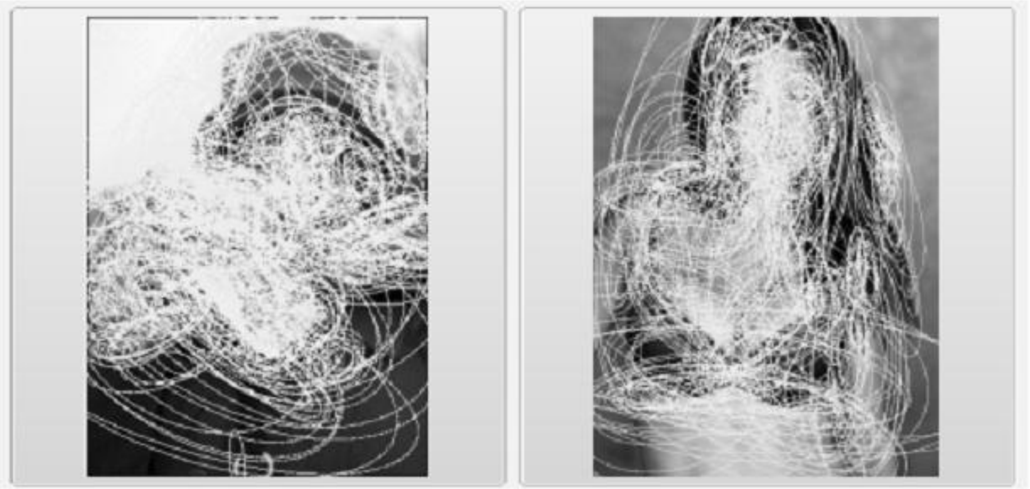


Figure 1. Interest points detected by Hessian-Affine detector

We start with the pre-processing step. The image features are extracted by the Scale-invariant feature transform method (SIFT [13]) which is locally based on the appear¬ance of the object at particular interest points, invariant to image scale, rotation and also robust to changes in illumination, noise, occlusion. Local descriptors in an im¬age are also computed in two stages: we first detect the interest points in the image.

These points are either maximums of Laplace of Gaussian, or 3D local extremes of Difference of Gaussian [14], or the points detected by a Hessian-Affine detector [16]. Figure 1 shows some interest points detected by a Hessian-Affine detector.

The descriptor of interest points is then computed on gray level gradient of the region around the point. The scalable invariant feature transform descriptor, SIFT [13] is often preferred. Each SIFT descriptor is a 128-dimensional vector. An example of SIFT is shown in figure 2.

The next step is to form visual words from the local descriptors computed in the previous step. Most of works perform a k-means [15] on descriptors and take the averages of each cluster as visual word [1, 5, 12]. After building the visual vocabulary, each descriptor is assigned to the nearest cluster. For this ends, we com-pute, in R128, distances from each descriptor to the representatives of previously defined clusters. Thus, an image is characterized by the frequency of its descriptors and the image corpora will be represented in the form of a contingency table crossing images and clusters (visual words). Figure 3 describes the BoVW model for representing images.

The pre-processing steps bring out datasets with a very large number of dimensions (e.g. 3000 visual words with many input features with each one containing only a small amount of information).

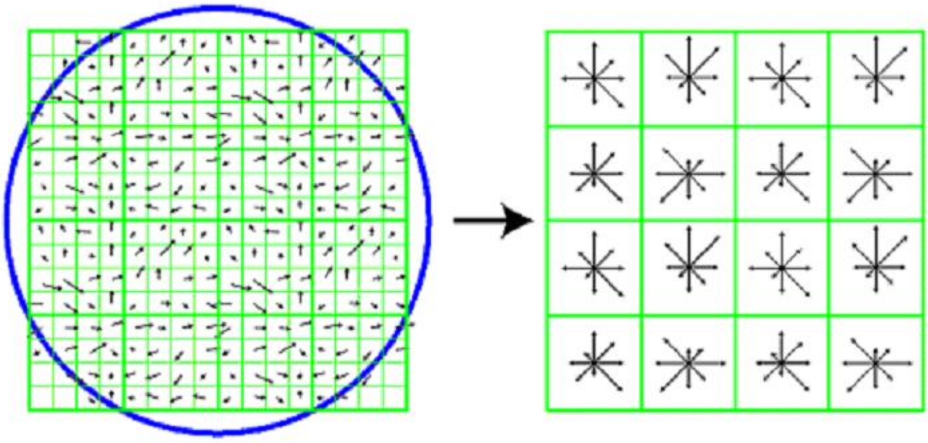


Figure 2. SIFT descriptor computed from the region around the interpret point (the circle): gradient of the image (left), descriptor of the interest point (right)

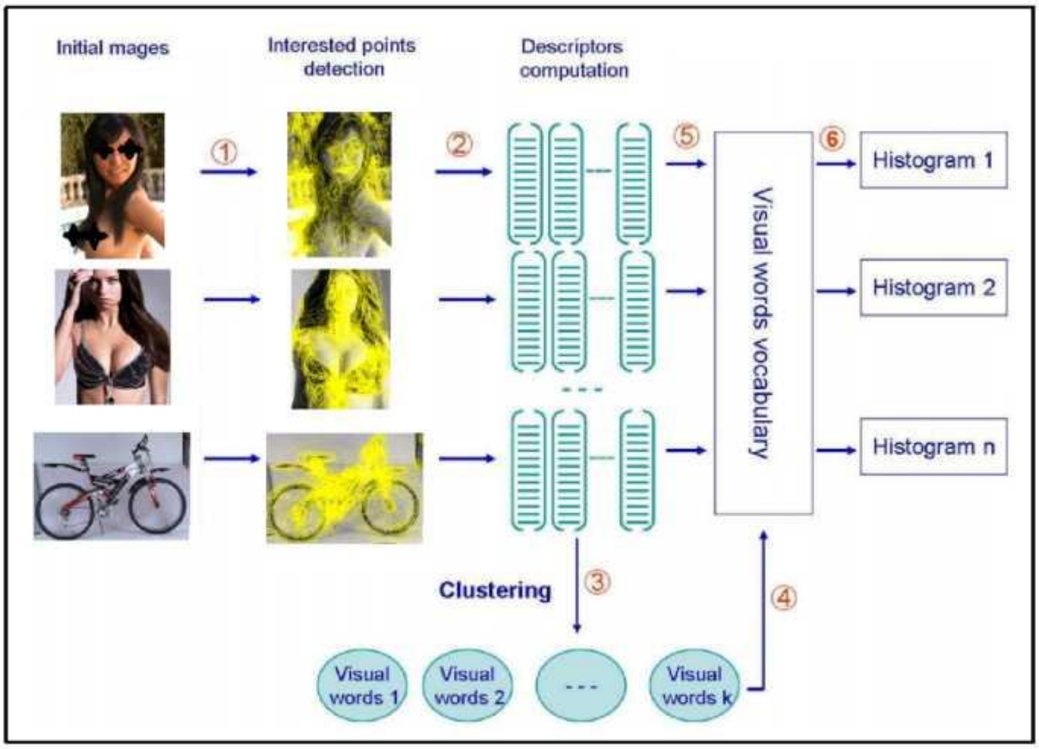


Figure 3. Bag-of-Visual-Words model for representing images

3. ARCX4 OF RANDOM MULTINOMIAL NAIVE BAYES (ARCX4-RMNB)

The image representation as bags of visual words is considered as a text categorization tasks using the bag-of-words model. There are many machine learning algorithms for this problem [20]. Due to the simplicity, the multinomial naive Bayes algorithm [11] is frequently used for text classification tasks.

3.1. Multinomial naive Bayes (MNB)

Let us considered a classification task using the training set of documents be denoted by T with N different words and C classes. The MNB algorithm assigns a test document t_i to the class be having the highest probability $P(c|t_i)$ using Bayes' rule, is given by:

$$P(c|t_i) = \frac{P(c)P(t_i|c)}{P(t_i)} \qquad c \in C \tag{1}$$

The class prior probability $P(c)$ is easy to be estimated by dividing the number of documents of class c by the training set size. Note that $P(t_i)$ is unchanged while computing class probabilities for every class. Therefore, $P(t_i)$ can be ignored without any change in results. The probability of obtaining a document like ti in class c, $P(t_i \mid c)$ is estimated as follows:

$$P(t_i \mid c) = (\sum_n f_{ni})! \prod_n \frac{P(w_n \mid c)^{f_{ni}}}{f_{ni}!} \tag{2}$$

where f_{ni} is the frequency of word w_n in test document t_i and the probability of word w_n given class c, $P(w_n \mid c)$ is estimated by dividing the frequency of word w_n in the training documents of class c by the total count of every words in the training documents belonging to class c. Due to the computational expensive terms $(\sum_n f_{ni})!$ and $\prod_n f_{ni}!$ in equation 2, $P(t_i \mid c)$ is estimated by suppressing them without any change in results.

MNB algorithm is computationally very efficient and easy to implement. It achieves good results for text classification tasks [11]. However, this algorithm does not give good performances compared with others ones, e.g. SVM [21] while dealing with very-high-dimensional datasets (many input features with each one containing only a small amount of information), typically the image representation using BoVW models.

3.2. Arcx4 of multinomial naive Bayes (Arcx4-rMNB)

Our investigation aims at improving classification results of MNB for very-high-dimensional cases. We propose a new learning algorithm, called Arcx4 of random multinomial naive Bayes (Arcx4-rMNB) that is suited for classifying this kind of data. Our Arcx4-rMNB algorithm uses Arcx4 approach [2] to constructs sequentially k random multinomial naive Bayes (rMNB) so that each weak classifier (i.e. rMNB) tries to focus mostly on the errors produced by the previous ones.

Since the nineties the machine learning community studies how to combine multiple weak classifiers into a strong ensemble-based model that is more accurate than a single one. The purpose of ensemble classifiers is to reduce the variance and/or the bias in learning algorithms. Bias is the systematic error term (independent of the learning sample) and variance is the error due to the variability of the model with respect to the learning sample randomness. Boosting (AdaBoost) proposed by [9] and Arcing (Arcx4) proposed by [2] aim at simultaneously reducing the bias and the variance.

Arcx4 algorithm calls repeatedly a given weak or base learning (e.g. decision trees, MNB) algorithm k times so that each iterative step concentrates mostly on the errors produced by the previous steps. For achieving this goal, it needs to maintain a distribution weights over the training datapoints. Initially, all weights are set equally and at each iterative step the weights of misclassified datapoints are increased so that the weak learner is forced to focus on the hard examples in the training set. The final model uses a majority vote of k weak classifiers.

- Initialize: distribution

$$d^0(i) = \frac{1}{trainsize}$$

- For k^{th} step

1. sampling S^k is based on d^{k-1}
2. training $rMNB^k$ model from S^k, using random subset features ($N' \approx \sqrt{N}$ original features) to estimate class probabilities.
3. updating distribution

$$d^k(i) = \frac{(1+m(i)^4)}{\sum_{p=1}^{trainsize}(1+m(p)^4)}$$

where $m(i)$ is the number of misclassifications of the i^th datapoint by

$rMNB^1, rMNB^2, \ldots, rMNB^k$

- Prediction of a new datapoint x: majority vote among outputs

$rMNB^1(x), rMNB^2(x), \ldots, rMNB^k(x)$

Alternately, our Arcx4-rMNB uses rMNB algorithm as a weak classifier at each iterative step. Furthermore, we propose to use random subset features ($N' \approx \sqrt{N}$ original features) to estimate class probabilities of MNB. This idea increases noise robustness of MNB [3]. Therefore, the ArcX4-rMNB can deal with classification tasks be having many input features with each one containing only a small amount of information.

4. NUMERICAL TEST RESULTS

We are interested in the accuracy of the new proposal (BoVW and Arcx4-rMNB) for pornographic images classification system. We here report the comparisons of the performance obtained by Arcx4-rMNB and the state-of-the-art algorithms, including MNB, SVM [21], decision tree C4.5 [17], AdaBoost of C4.5 (AdaBoost-C4.5 [9]) and random forests (RF-CART [3]). In order to evaluate performance for classification tasks, we have implemented Arcx4-rMNB and MNB in C/C++. We also use the highly efficient standard SVM algorithm LibSVM [4]. The rest are implemented in Weka library [23].

4.1. Experimental results

We do setup experiment with two real datasets[1] for comparative studies. We have created the first dataset be having 1414 images and the second large one be having 14971 images in two classes *(porno* and *non - porn),* described in table 1. Non-pornographic images are collected from the video frames of TF1 Channel in France, the advertisements of underwear (e.g. Triumph) and the auto exposition. Some examples of images are represented in figure 4.

[1] Datasets available to Researchers only upon request by email (dtnghi@cit.ctu.edu.vn)

Table 1. Description of two pornographic image datasets

ID	Datasets	#porno images	#non-porno images
1	Small base	484	930
2	Large base	6944	8027

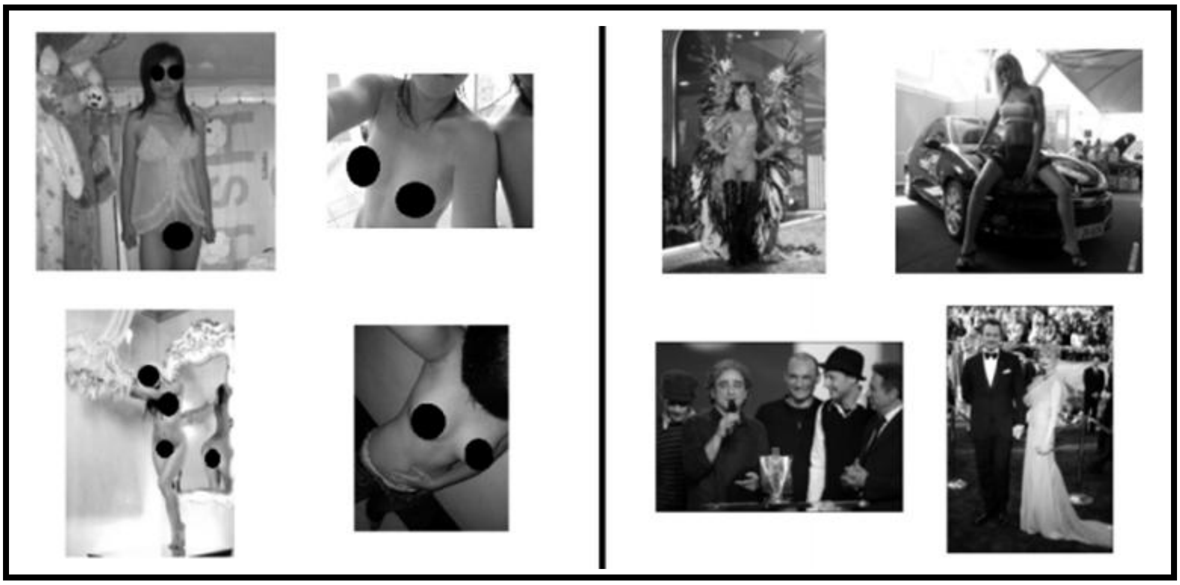


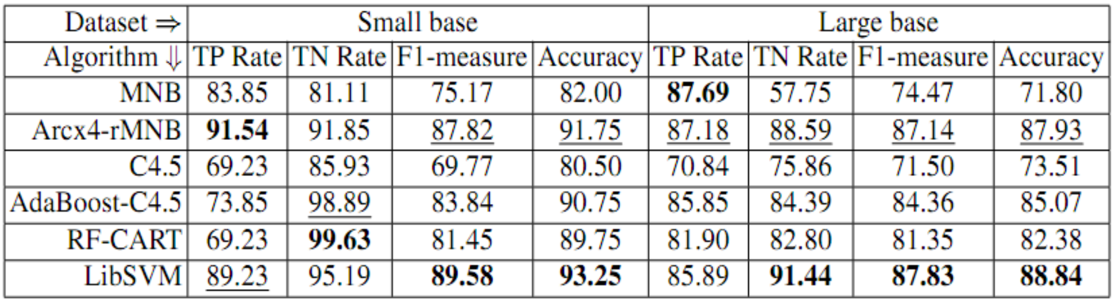
Figure 4. Examples of pornographic (left) and non-pornographic (right) images

Due to the BoVW model for representing images, we used also the Hessian- Affine SIFT detector implemented by Mikolajczyk and Schmid [16] to extract local descriptors which are grouped into 3000 clusters (visual words) with k-means algorithm. After building the visual vocabulary, the representation of the image that we use for classification is the (BoVW) model which is constructed from the local descriptors and the counting of the occurrence of visual words in a histogram like fashion. The pre-processing step brings out two datasets (table) with 1414 dat- apoints and 14971 datapoints in 3000 dimensions respectively. We remark that we tried to vary the number of clusters (visual words from 1000 to 50000) for finding the best experimental results. And then, we obtained accurate models with 3000 visual words and it seems that the results are unchanged while increasing the number of visual words over 3000.

The performance of the classification algorithms is compared in terms of correctness rate for each class (true positive - TP, true negative - TN), F1-measure and accuracy [18]. The precision for a class (i.e. *porno)* is the number of datapoints correctly labeled as belonging to the class divided by the total number of datapoints labeled as belonging to the class. The recall for a class is the number of datapoints correctly labeled as belonging to the class divided by the total number of elements that actually belong to the class. The F1-measure is a synthesis of the precision and the recall, which is defined as the harmonic mean of these both quantities. The accuracy is the number of datapoints correctly labeled divided by the total number of elements of the dataset.

We propose to use the test protocol called hold-out as follows: The dataset is randomly partitioned into a training set (be having 2/3 dataset) and a testing set (the remaining). We used the training set to tune the parameters of the algorithms including Arcx4-rMNB, AdaBoost-C4.5, RF-CART, LibSVM for obtaining a good accuracy in the learning phase. Then the obtained model is evaluated on the testing set. The process is then repeated 3 times. The results are then averaged to produce the final result. We tried to use different kernel functions of the SVM algorithm, including a polynomial function of degree d , a RBF (RBF kernel of two datapoints x_i , x_j , $K[i, j] = exp(-\gamma\| x_i - x_j\|^2))$. Thus, the SVM algorithm using the RBF kernel (with $\gamma = 0.0002$) gives the best results. Arcx4-rMNB learns 200 weak classifiers (rMNB) using randomly 300 dimensions for estimating class probabilities. RF-CART builds 200 trees using randomly 300 dimensions for performing multivariate node splitting. AdaBoost-C4.5 also performs classification tasks using 200 trees.

Table 2. Classification results on two pornography image datasets



Dataset ⇒	Small base				Large base			
Algorithm ⇓	TP Rate	TN Rate	F1-measure	Accuracy	TP Rate	TN Rate	F1-measure	Accuracy
MNB	83.85	81.11	75.17	82.00	**87.69**	57.75	74.47	71.80
Arcx4-rMNB	**91.54**	91.85	87.82	91.75	87.18	88.59	87.14	87.93
C4.5	69.23	85.93	69.77	80.50	70.84	75.86	71.50	73.51
AdaBoost-C4.5	73.85	98.89	83.84	90.75	85.85	84.39	84.36	85.07
RF-CART	69.23	**99.63**	81.45	89.75	81.90	82.80	81.35	82.38
LibSVM	89.23	95.19	**89.58**	**93.25**	85.89	**91.44**	**87.83**	**88.84**

The main results are showed in table 2. The best results are bold faces and the second ones are underlined. The plot charts in figure 5 and figure 6 visualize the classification results in terms of precision, recall, F1, accuracy, TP rate and TN rate.

In comparison of classification results obtained by Arcx4-rMNB facing the state- of-the-art algorithms, our Arcx4-rMNB outperforms MNB, C4.5 [17], RF-CART [3], AdaBoost of C4.5 [9], in terms of correctness rate of $porno$ detection, F1 measure and accuracy.

For a more detailed assessment of the performance obtained by Arcx4-rMNB facing LibSVM. Our Arcx4-rMNB detects more accurate of pornographic images but makes more false alarm than LibSVM. Therefore, our Arcx4-rMNB outperforms LibSVM in terms of correctness rate of $porno$ but LibSVM gives best results in terms of correctness rate of $non-porno$, F1-measure and accuracy.

With these classification results, we believe that the Arcx4-rMNB achieves good performances while comparing with the state-of-the-art classifier algorithms.

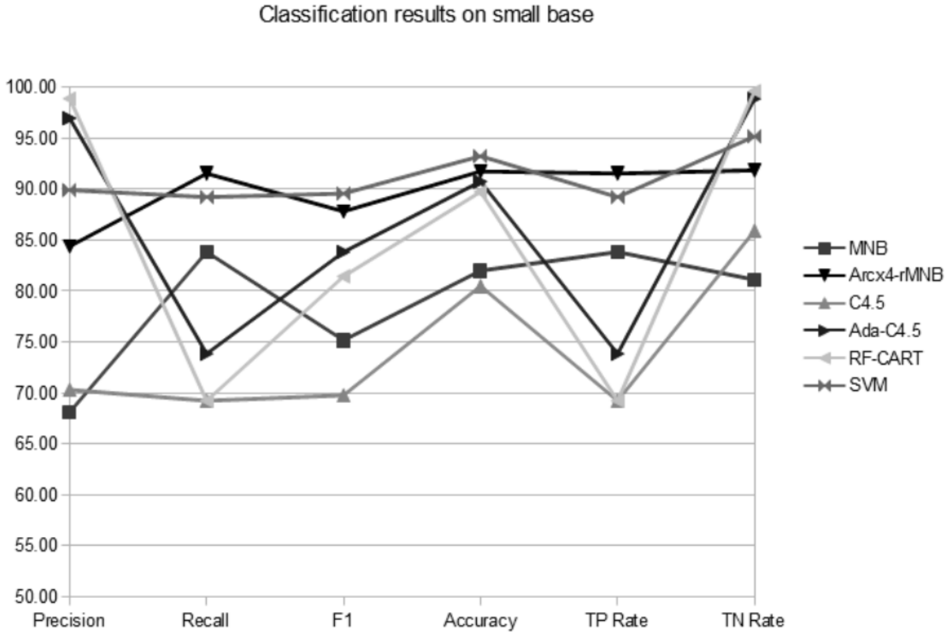


Figure 5. Classification results on small dataset

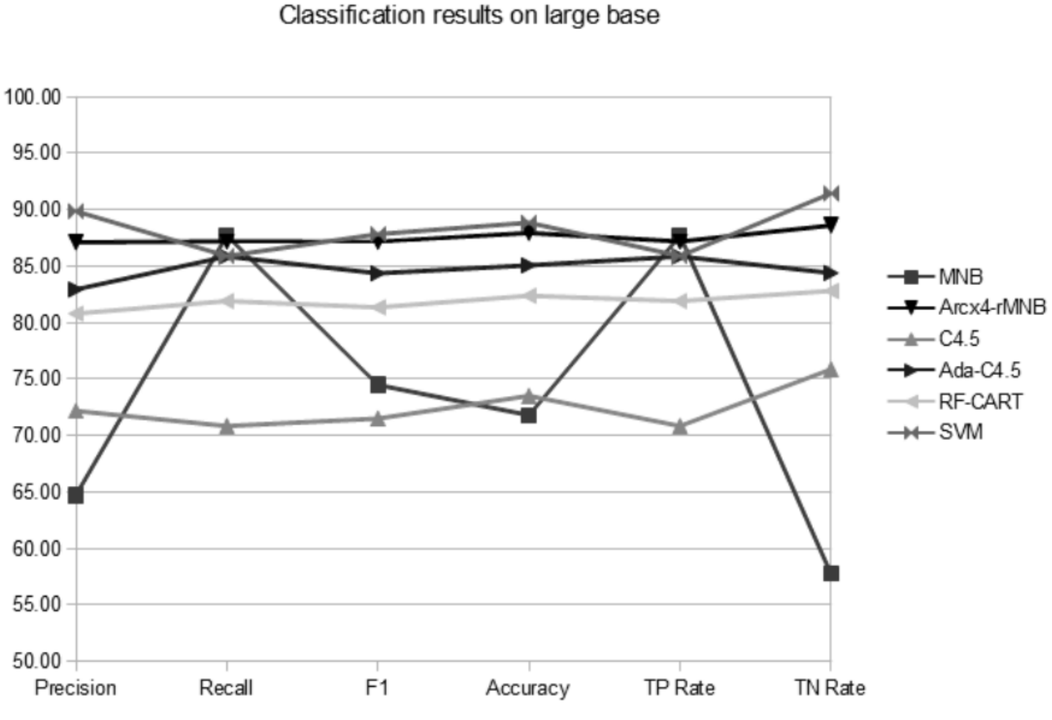


Figure 6. Classification results on large dataset

5. CONCLUSION AND FUTURE WORKS

We presented a novel approach that achieves high performances for classification tasks of pornographic images. The main idea is inspired from the bag-of-visual- words (BoVW) model and Arcx4 of random multinomial naive Bayes (Arcx4- rMNB) algorithm. The representation of the image is constructed from the local descriptors obtained by the Hessian-Affine SIFT detector and the counting of the occurrence of visual words. The representation of the image based on the BoVW model brings out very-high-dimensional datasets. And then, we propose a new learning algorithm Arcx4-rMNB that is suited for classifying very-high-dimensional datasets, including the pornographic image representation using BoVW models. The numerical test results on two real datasets of images showed that our proposal has achieved an accuracy of 91.75% for a small dataset and 87.93% for other large one. Our Arcx4-rMNB algorithm outperforms the state-of-the-art algorithms including decision tree C4.5 [17], random forest of CART (RF-CART [3]), AdaBoost of C4.5 [9]. Our Arcx4-rMNB detects more accurate of pornographic images but makes more false alarm than SVM [21]. These results showed that our proposal (the BoVW model and the Arcx4-rMNB algorithm) achieves a high accurate compared with related works [7, 8, 5, 6, 10, 24].

In the future, we intend to apply this approach into pornographic video classification tasks. In addition, our proposal is very efficient and it can be parallelized. A parallel implementation that exploits the multicore processors can greatly speed up the learning tasks.

Acknowledgements. The author would like to thank to Minh-Trung Nguyen and Nguyen-Khang Pham for this help with preparing the collections of images.

REFERENCES

1. Bosch, A., Zisserman, A., Munoz, X. - Scene classification via pLSA. In: Proceedings of the European Conference on Computer Vision, 2006, pp. 517-530.
2. Breiman L. - Arcing classifiers, The annals of statistics **26** (3) (1998) 801-849.
3. Breiman L. - Random forests, Machine Learning **45** (1) (2001) 5-32.
4. Chang C. C., Lin, C. J. - LIBSVM - a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm
5. Deselaers T., Pimenidis L., Ney H. - Bag-of-visual-words models for adult image classification and filtering, In: Proceeding of The 19th International Conference on Pattern Recognition, 2008, pp. 1-4.
6. Duan L., Cui G., Gao W., Zhang H. - Adult image detection method base-on skin color model and support vector machine, In: Proceeding of The 5th Asian Conference on Computer Vision, 2002, pp. 797-800.
7. Fleck M., Forsyth D., Bregler C. - Finding naked people, In: Proceedings of the

European Conference on Computer Vision **2** (1996) 592-602.

8. Forsyth D., Fleck M. - Identifying nude pictures. In: Proceedings of the IEEE Workshop on the Applications of Computer Vision, 1996, pp. 103-108.

9. Freund Y., Schapire R. - A decision-theoretic generalization of on-line learning and an application to boosting, In: Computational Learning Theory: Proceedings of the Second European Conference, 1995, pp. 23-37.

10. Jeong C., Kim J., Hong K. - Appearance-based nude image detection, In: Proceedings of The 17th International Conference on Pattern Recognition, 2004, pp. 467-470.

11. Lewis D., Gale W. - A sequential algorithm for training text classifiers, In: Proceedings of SIGIR, 1994.

12. Lopes A., Avila S., Peixoto A., Oliveira R., Coelho M., Araiijo A. - Nude detection in video using bag-of-visual-feature, In: Proceedings of The 22th Brazilian Symposium on Computer Graphics and Image Processing, 2009, pp. 224-231.

13. Lowe D. - Object recognition from local scale invariant features. In: Proceedings of the 7th International Conference on Computer Vision, 1999, pp. 1150-1157.

14. Lowe D. - Distinctive image features from scale invariant keypoints. International Journal of Computer Vision, 2004, pp. 91-110.

15. MacQueen J. - Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press 1, 1967, pp. 281-297.

16. Mikolajczyk K., Schmid C. - Scale and affine invariant interest point detectors, International Journal of Computer Vision **60** (1) (2004) 63-86.)

17. Quinlan J. R. - C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

18. van Rijsbergen C. V. - Information Retrieval, Butterworth, 1979.

19. Schettini R., Brambilla C., Cusano C., Ciocca G. - On the detection of pornographic digital images, In: Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, 2003, pp. 2105-2113.

20. Sebastiani F. - Machine learning in automated text categorization, ACM Computing Surveys **34** (1) (1999) 1-47.

21. Vapnik V. - The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

22. Wang Y., Wang W., Gao W. - Research on the discrimination of pornographic and bikini images, In: Proceedings of the Seventh IEEE International Symposium on Multimedia, 2005, pp. 558-564.

23. Witten I., Frank E. - Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.

24. Zheng H., Daoudi M. - Blocking adult images based on statistical skin detection, Electronic Letters on Computer Vision and Image Analysis **4** (2) (2004) 1-14.

Corresponding author:

Thanh-Nghi Do,

Institut Telecom; Telecom Bretagne
UMR CNRS 3192 Lab-STICC
Universite europeenne de Bretagne, France

Email: *tn.do@telecom-bretagne.eu* - *dtnghi@cit.ctu.edu.vn*