IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 6, December, 2023, pp. 1679-1687
Journal homepage: http://iieta.org/journals/isi

Enhancing Identity Document Classification in KYC Processes: An Evaluation of the Bag-of-Visual-Words Model and Segmentation Impact

Candy Lee[1*], Iman Herwidiana Kartowisastro[1,2]

[1] Computer Science Department, Bina Nusantara University, Jakarta 11530, Indonesia
[2] Computer Engineering Department, Faculty of Engineering, Bina Nusantara University, Jakarta 11530, Indonesia

Corresponding Author Email: candy.lee@binus.ac.id

https://doi.org/10.18280/isi.280626

Received: 28 September 2023
Revised: 27 November 2023
Accepted: 7 December 2023
Available online: 23 December 2023

Keywords:
Bag of Visual Words (BoVW), identity document classification, k-means, Scale Invariant Feature Transform (SIFT), Support Vector Machine (SVM), vector quantization

ABSTRACT

The growth of online services, such as financial services, travel agencies, and e-government, has emphasized the importance of an efficient Know Your Customer (KYC) process. Efficient identity verification and document classification are crucial for KYC, such as ensuring the alignment of submitted identity documents with requirements, categorizing them accurately, and verifying their completeness within the KYC process. This article proposes the utilization of the bag-of-visual words (BoVW) model, which combines SIFT, k-means, and SVM techniques, to achieve accurate identity document classification without relying on geometry transformations. We observed that while segmentation significantly enhances accuracy during testing by eliminating irrelevant parts, its impact on the training phase appears to result in a drop in the model's performance. This drop in performance might be associated with segmentation during the training phase, where the removal of irrelevant parts might have caused the algorithm to have difficulty in identifying which features to disregard within the samples. This also implies that introducing imperfections such as blurred and low brightness samples into training dataset could potentially enhance the classification model. To test the theory, we compiled a dataset consisting of 8,400 samples, divided into 20 classes. This single compiled dataset was then used to generate three different kinds of datasets: USGM (an unsegmented dataset), SGM (a segmented dataset), and SGM2 (a segmented dataset where the subject of interest is clearly visible in the samples, serving as the training dataset). Three different testing is used: same-variant, cross-variant, and k-fold cross-validation. Our model demonstrates an average accuracy up to 97.2%, which remains relatively consistent across different types of testing.

1. INTRODUCTION

The expan the growth of online services, such as financial technology (FinTech) [1], internet banking, travel websites, and government online services [2], and more, has emphasized the importance of an efficient Know Your Customer (KYC) process. The KYC process itself is a crucial process where companies assess their potential users to ensure that their potential users can be trusted [3].

In every KYC process, a potential user or customer is expected to submit proof of their identity to the service provider [1], typically in the form of a scanned document or a photo of an official identification document. An identification document typically contains details like name, address, place and date of birth, and nationality. Some identity documents may also include additional information, such as religious affiliation [4].

From here, we can infer the necessary steps. The company must identify the type of identity document received from the user, authenticate its validity, and subsequently extract and digitize the information for further processing within the KYC procedure.

Accurate determination of the type and country of origin of identity documents can significantly aid companies in streamlining the initial phases of the KYC process. This involves ensuring that submitted identity documents align with the specified requirements and verifying their completeness. It also aids in providing a context for authenticating these documents and helps the system determine the anticipated language for Optical Character Recognition (OCR). Failure to do so could lead to unnecessary frustration for both users and service providers, leading to iterative exchanges in the KYC process, thereby causing delays or even limiting the company's capacity for customer acquisition.

In our investigation into classification approaches for identity documents, we discovered that the majority of methods rely extensively on visual features. Notably, one approach employs the Bag-of-Visual-Words (BoVW) technique [5], utilizing the SURF algorithm as a feature descriptor. Another method involves the combined usage of SURF, direct matching, and random sample consensus (RANSAC) [6].

In comparison to SIFT, SURF demands relatively fewer computational resources while producing fewer descriptors, albeit with trade-offs in lower accuracy [7, 8]. This reduction in accuracy can be attributed to several factors, one them being number of feature descriptor generated by SURF algorithm is fewer compared to SIFT. Given that identity documents inherently possess fewer distinctive features, a reduced availability of descriptors could lead to decreased model accuracy. We assert that conducting the classification process on the server side, leveraging ample resources that can be scaled as needed, mitigates the concern of conserving computational resources. Additionally, SIFT demonstrates greater resilience to noise and changes in lighting conditions [7], characteristics aligning with scenarios where customers might submit identity documents captured under various uncontrolled lighting conditions and diverse camera technologies.

Some methods attempt to integrate visual features with other features. For example, using a histogram of gradients and color of the document, which is visual features along with the document's spatial information as a feature [9]. However, another study argues that, in the case of identity document classification, the availability of distinctive features extractable for classification purposes is limited. For instance, many identity documents share the same layout regardless of their type or issuing country [6]. Visual features can also be used along with textual features; this approach is demonstrated using SIFT to detect visual features and OCR to extract textual information [10]. However, OCR requires prior knowledge of the expected language within the document, as highlighted in the study.

Lastly, neural network based method such as CNN [11] have also been suggested to classify identity documents. Despite achieving relatively high accuracy, these methods demand extensive computational resources and substantial datasets to yield high-performing models.

Our focus on delving deeper into the BoVW method is motivated by several observations within the context of classifying identity documents for KYC processes. Firstly, the types of photos users submit are often captured under various uncontrolled lighting conditions and diverse camera technologies. Secondly, in scenarios involving sensitive information, such as processing identity documents, it is preferable to handle the process as much as possible within the company itself, rather than outsourcing, to prevent data leaks. However, the availability of samples for model training is limited. Hence, an ideal algorithm for this purpose should not require a large dataset for accurate predictions.

Aside from method selection, our investigation indicates that improving model performance often involves tuning the dataset used for training and adjusting algorithm hyperparameters. Therefore, this study explores strategies for dataset compilation, focusing on tuning hyperparameters, such as the cluster size in k-means within the proposed method. We perceive clustering descriptors into visual words as a crucial aspect of the BoVW, emphasizing the importance of finding an optimized cluster size to achieve an efficient classification model. To summarize, this study aims to explore dataset compilation strategies and the tuning of critical hyperparameters.

In addition, this study also aims to investigate how pre-processing such as segmentation, geometric transformation, and orientation affect the classification results. All pre-processing is a whole field by itself and will add a layer of complexity to the model. Pre-processing constitutes a distinct field on its own and introduces an additional layer of complexity to the model. For instance, a study has shown that imperfections in segmentation may potentially cause a model's performance to drop [5]. Fixing perspective poses a challenge by itself. First, the segmentation has to be precise; secondly, the orientation fixing might fail, causing the samples to be rejected even before the classification process [6].

We believe that our study can offer a novel perspective on dataset compilation strategies, the inclusion of pre-processing in the workflow, and how tuning cluster size can enhance the classification of identity documents.

The rest of this paper is structured as follows: In Section 2, we present related works that we have studied. In Section 3, we describe the methodology used in this study. Discussion of the results of the research that has been done is presented in Section 4. Finally, in Section 5, we draw conclusions and suggest how this research can be developed further.

2. RELATED WORKS

We observed that the research paper on the topic of identity document classification is quite limited. This can be caused by some restriction [12] of researchers in collecting identity documents as a research dataset. Therefore, we have also tried to study other classification methods such as image classification in general and text document classification, apart from the classification methods developed to classify identity documents.

In the context of image classification, a method was proposed by Karim and Sameer [13] that utilized a bag of visual words (BoVW) with scale-invariant feature transform (SIFT) as a descriptor and k-nearest neighbor (k-NN) for classification in vehicle image classification. Despite a relatively small number of datasets, the researchers achieved good accuracy. Mittal and Saraswat [14] proposed a method for classifying histopathological images using BoVW with SIFT as a descriptor, GSA as a clustering method, and SVM for classification. The study revealed that GSA-based clustering outperformed k-means in the descriptors clustering stage, leading to improved results for histopathological cases. In the research conducted by Gao and Lee [15] to recognize car manufacturers and models from video, they utilized SIFT to detect keypoints on the front of a car. The frontal image was extracted through frame analysis, and SIFT was applied to identify keypoints on the extracted image. These keypoints were then matched with a database of car frontal images or a planar database using nearest neighbor search (NNS). Sarwar et al. [16] proposed a method to enhance the effectiveness of the Bag-of-Words model (BoW) for content-based image retrieval. They incorporated two characteristics, local intensity order pattern (LIOP) and local binary pattern variance (LBPV), which could be used separately or together to create vocabulary dictionaries. The study demonstrated that utilizing LIOP and LBPV separately improved recall performance, while combining them into a larger vocabulary increased accuracy or precision. Chaganti et al. [17] introduced CNN-SVM as a solution to the problem of image classification with a relatively large number of samples. The study highlighted that while SVM with a small number of samples could produce fairly good accuracy, its performance declined when faced with larger datasets. Therefore, the integration of a neural network such as a convolutional neural network (CNN) was found to significantly improve the classification accuracy. Chow and Reyes-Aldasoro [18] proposed a classification method for gemstone images based on color histograms and random forest. The researchers argued that while certain gemstones could be identified by both shape and color, others, such as Emerald and Tsavorite, could only be distinguished by color. Thus, a color-based classification approach was deemed reliable for gemstone classification.

In the case of document classification, Tensmeyer and Martinez [19] proposed the CNN method to classify text-based documents. In this study, the researchers experimented with various methods to preprocess the input, including resizing, mirroring, cropping, and shear transformation, among others. The study concluded that shear transformation and large input images contributed the most to achieving better performance. The proposed method was also observed to learn layout features such as graphics, type-set text, handwriting, etc., from the document without prior information provided to CNN. Zhao and Mao [20] proposed an improvement for the BoW method, named the Fuzzy Bag-of-Word Clusters (FBoWC) model. FBoWC integrated fuzzy mapping into bag-of-visual words (BoVW) and used a cluster of words instead of individual words to build document representation. The study introduced three different variants of FBoWC based on the similarity features of the word cluster: FBoWCmean, FBoWCmax, and FBoWCmin. The study concluded that FBoWC could reduce feature redundancy and improve discriminant features. Yao et al. [21] proposed text graph convolutional networks (text-GCN) to solve document classification problems. However, on a dataset with minimal text information, the authors found that the proposed method struggled to find the relations between words to perform the classification task. Asim et al. [22] proposed a two-stream analysis that combined textual and visual features for document classification tasks. In the first stream, the study proposed a method to extract textual information using OCR and rank them based on their ability to discriminate document images. In the second stream, InceptionV3 was used to extract visual features from the sample document. Finally, the textual and visual streams were concatenated using an average ensembling method. The study concluded that the proposed method could detect more discriminant features by combining textual and visual features.

In identity document classification, De Las Heras et al. [5] suggested that by adding imperfect samples, such as blurry images, during the training phase, accuracy could be improved. The study also found that spatial information did not significantly improve accuracy. The author of the paper argued that this lack of improvement could be attributed to imperfections in the segmentation process. Simon et al. [9] proposed a combination of three different methods to classify identity documents: Histogram of Gradients (HoG), the color of the document, and spatial pyramids with a depth of three (SP3). The method only required one sample to train their model. Their study suggested that, by adding spatial information, performance could be improved. The study also evaluated the OCR approach to classify identity documents and found that the main reason why the OCR approach did not perform well was the unknown language and font used in different document samples. Sicre et al. [11] compared several methods for the classification of identity documents. The study showed that by utilizing CNN as a descriptor, combining it with a vector of locally aggregated descriptors (VLAD) for descriptor clustering into visual words, and utilizing SVM for classification, a high-accuracy model could be achieved. Awal et al. [6] conducted a study on performing identity document classification without the need for a training phase. The authors explored a methodology where a document model was created based on a single reference image. In the prediction process, the document under consideration was compared against all the models stored in the database. Once a matching model was found, a more intricate analysis was carried out to determine whether the matched model should be accepted or rejected. Khandan [10] proposed a method that combined SIFT and OCR for identity document classification. The study aimed to develop a method to classify identity documents with a confidence level for each match, which was determined by the number of matches in the SIFT model during each classification task. However, it should be noted that the dataset used in the study contained samples from a single country, implying the presence of a single language. In another study that aimed to detect fabricated identity documents [23], a novel descriptor, grid color connected components descriptor (Grid-3CD), was proposed. Grid-3CD could be used to extract information such as the color, position, and shape of an image from the sample.

The approach to document classification can be categorized into three types: analyzing visual features, utilizing textual information, and analyzing document layout or spatial information. In the field of general image classification, where text is absent, visual features have gained popularity. Visual features can be further divided into gradient-based approaches [13-16] and color-based approaches [18]. On the other hand, when it comes to document classification, the analysis of textual information emerges as the most common approach [20, 21]. Some studies have also shown that incorporating visual features can enhance text-based document classification [22]. In the context of identity document classification, visual features often serve as the basis for classification methods [5, 6], although some studies explore the combination of multiple types of information such as visual features and textual information [10], or visual features and spatial information [5, 9]. CNN has been proposed as a classification method for all three categories examined in our study: general image classification [17], text document classification [19], and identity document classification [11]. It is generally observed that CNN achieves superior performance, but it requires a substantial amount of datasets to yield effective results.

One of the challenges encountered in identity document classification is the limited availability of distinctive features that can be extracted for classification purposes. For instance, many identity documents share the same layout regardless of their type or issuing country [6]. When considering a textual approach, two challenges arise. Firstly, detecting the area of the text within the document can be a task in itself [11, 23]. Secondly, language adds another layer of complexity, as the performance of technologies such as OCR can vary depending on the language used in the identity document [11]. As a result, visual features emerge as one of the most viable options for performing this task. Despite their minimal or subtle differences, they still possess enough discriminative capability to be effectively utilized in classification.

Regarding datasets, we discovered that the scarcity of research on identity documents compared to other document types can be attributed to the sensitive nature of these documents, which often contain confidential information. In certain studies, researchers addressed this challenge by collecting their own data [6, 9] or collaborating with companies [10, 11] to obtain datasets. However, this approach restricts access to the dataset by other researchers, limiting reproducibility. Therefore, we opted to utilize publicly available datasets intended for academic research, such as

MIDV-500 [12] and MIDV-2019 [2], in order to enhance the reproducibility of our study. The two datasets were selected because they contain samples of identity document photo captured under various conditions and using more than one type of smartphone.

3. METHODOLOGY

In this study, we utilized the Bag-of-Visual-Words (BoVW) method for identity document classification. The BoVW model incorporates SIFT, which is used to detect descriptors, k-means clustering for grouping these descriptors, and vector quantization for constructing a visual vocabulary. The resulting visual vocabulary is then utilized for classification into pre-defined classes using an SVM (see Section 3.2). Different k values in the k-means algorithm were explored to investigate the potential improvement in model performance. Three testing approaches were used: same-variant, cross-variant, and k-fold cross-validation (see Section 3.3). Accuracy was used as the metric to assess the performance. Figure 1 shows the methodology used in this experiment.

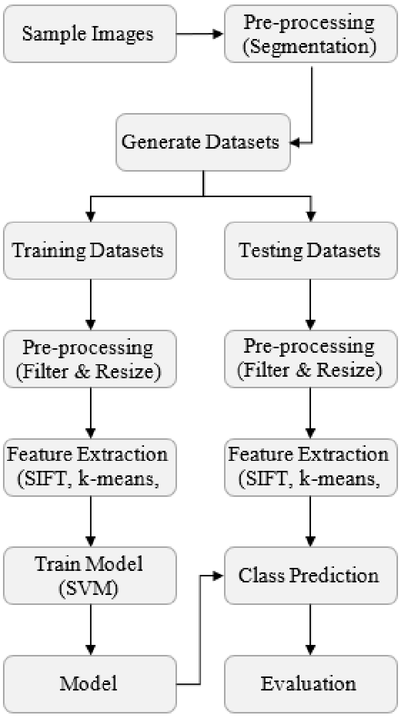

Sample Images
Pre-processing (Segmentation)
Generate Datasets
Training Datasets
Testing Datasets
Pre-processing (Filter & Resize)
Pre-processing (Filter & Resize)
Feature Extraction (SIFT, k-means,
Feature Extraction (SIFT, k-means,
Train Model (SVM)
Class Prediction
Model
Evaluation

Figure 1. Methodology

3.1 Dataset

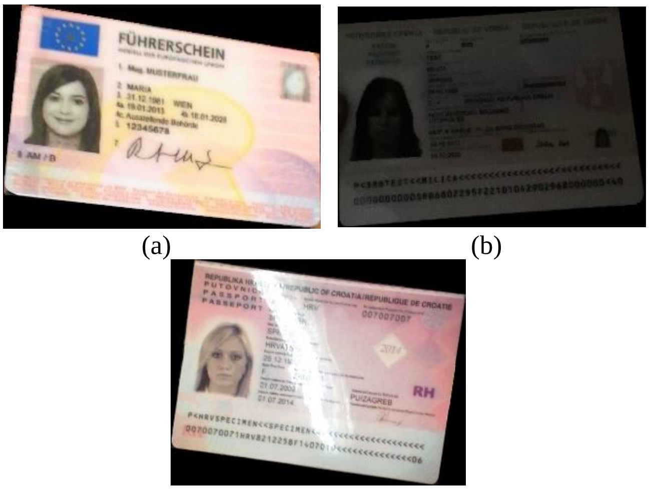
We combined the MIDV-500 [12] and MIDV-2019 [2] datasets together for this experiment. The two datasets were chosen based on two reasons, it was available for academic research and both datasets consist of identity document images captured under various conditions, using more than one type of smartphone. The MIDV-500 dataset consisted of identity documents categorized into five sub-categories: on a cluttered background, held in hand, on the keyboard, on the table, and with partial visibility. Similarly, the MIDV-2019 dataset included identity documents categorized into two sub-categories: distorted perspective and low-light conditions.

The combined dataset consisted of a total of 21,000 samples across 50 classes, but for this experiment, only 20 classes with 8,400 samples were included. To maintain focus, identity documents issued by a single country with a single document type were excluded. Additionally, if a newer version was available, the older version was omitted. As a result, 30 classes out of the initial 50 classes were excluded, leaving 20 classes remaining for the purpose of this experiment. Table 1 displays the list of classes utilized in our experiment.

Table 1. List of identity document classes

Class Name	Issuing Country	Document Type
aut.drvlic.new	Austria	Driving License
aut.id	Austria	ID Card
cze.id	Czech	ID Card
cze.passport	Czech	Passport
deu.drvlic.new	German	Driving License
deu.id.new	German	ID Card
deu.passport.new	German	Passport
esp.drvlic	Spain	Driving License
esp.id.new	Spain	ID Card
fin.drvlic	Finland	Driving License
fin.id	Finland	ID Card
hrv.drvlic	Croatia	Driving License
hrv.passport	Croatia	Passport
srb.id	Serbia	ID Card
srb.passport	Serbia	Passport
ukr.id	Ukraine	ID Card
ukr. passport	Ukraine	Passport
usa.bordercrossing	United States	Passport
usa.passportcard	United States	Passport
usa.ssn82	United States	ID Card

The datasets also included the coordinates indicating the location of the identity document within the samples, which were utilized in the segmentation process. Segmentation was performed without correcting distortions like geometry transformation or rotation. After removing the background, it was replaced with black color, and the resulting image was then cropped to fit the segmented part. As a result, each image or sample obtained from the segmentation process would have a different dimension. Figure 2 shows a selection of identity document samples after the application of segmentation.


(a) (b)

(c)

Figure 2. Examples of segmented identity documents: (a) Austrian driving license; (b) Serbian passport; (c) Croatian passport

Three datasets were generated: segmented (SGM), unsegmented (USGM), and segmented2 (SGM2). SGM consisted of samples of identity documents with segmentation, while USGM consisted of samples without segmentation. USGM served as a comparison dataset to SGM to observe if segmentation could improve accuracy. SGM2 served as a comparison dataset to SGM to assess if adding imperfections to the training dataset could enhance classification performance. These datasets were divided based on the clarity of the samples. Identity document images in the categories of clutter, hand, keyboard, and table were considered clear, where the identity documents could be seen clearly, and were grouped as the training dataset. The remaining categories, including partial, distorted, and low-light samples, were considered unclear and formed the testing dataset. Both SGM and USGM were split using the 80-20 strategy, where 80% of the samples were used for training and 20% for testing. In the case of SGM2, some samples had to be removed to maintain the dataset within the 80-20 strategy, resulting in SGM2 having a smaller sample size compared to SGM and USGM. From this point onward, the models trained with training datasets will be referred to with the suffix "-A" (e.g., SGM-A), the testing datasets with the suffix "-B" (e.g., SGM-B), and the unsplit/merged datasets without any suffix (e.g., SGM). Table 2 shows a list of our dataset variants.

Table 2. List of dataset variant

Alias	**Split Rule**	**Dataset Sizes**		
		Merged	**Train**	**Test**
SGM	Segmented	8,400	6,750	1,680
USGM	SGM without segmentation	8,400	6,720	1,680
SGM2	Based on the clarity of the samples	6,000	4,800	1,200

3.2 Training phase

During the training phase, all identity document samples were loaded along with their pre-defined classes. Two pre-processing steps were applied at this stage. The first step was to exclude identity document samples with low visibility, which was determined using Eq. (1). If the visibility score of a sample was below 0.1 or the identity document image's visibility was less than 10%, they were rejected and excluded from further processing. The second pre-processing step involved resizing the images to a proportional width of 320 pixels.

$$Visibility = 1 - \frac{\text{Numbers of black pixel}}{\text{Total pixel}} \tag{1}$$

It is important to note that since pre-defined coordinates were used to segment identity document images, we generated a list of images where the visibility score was below 0.1. This list is used in every pre-processing step to exclude images from further processing.

SIFT was utilized to detect descriptors from the identity document samples. This algorithm analyzes the difference of Gaussian (DoG), as defined in Eq. (2):

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) \tag{2}$$

where, $G(x, y, \sigma)$ represents a Gaussian function with changing scale, σ represents the scale variable of the Gaussian function, x represents the horizontal coordinate within the Gaussian window, y represents the vertical coordinate within the Gaussian window.

The equation for the Gaussian function can be defined as Eq. (3):

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{x^2+y^2}{2\sigma^2}\right] \tag{3}$$

The next step involved applying the k-means algorithm to cluster the descriptors extracted from the image samples. This clustering process resulted in the generation of clusters, where each cluster represents a group of descriptors with similar characteristics. These clusters are commonly referred to as visual words or codebook entries. To explore the impact of different cluster sizes on the model performance, four distinct values were selected: 96, 128, 160, and 192. The four distinct values were selected by dividing 320, the maximum number of pixels in width that we have decided, then incrementally decrease or increase the amount by 32.

k-means clustering started by initializing centroids through randomly selecting points from the dataset. For each data point, the Euclidean distance, as defined in Eq. (4), was calculated between the data point and each of the k centroids. The data point was then assigned to the cluster associated with the nearest centroid. Subsequently, new centroids were determined for each cluster using Eq. (5). The process of recalculating distances and reassigning cluster centroids was repeated until the centroids no longer changed significantly. To reduce the complexity of the clustered visual words and achieve a more efficient representation, vector quantization was used. This quantization step assigned each local feature (SIFT descriptor) to the nearest visual word in the visual vocabulary generated by k-means clustering. As a result, the descriptor was transformed into a collection of visual words, and the histogram of visual words was generated, representing the frequency distribution of these visual words. This histogram served as a global representation of the trained samples, capturing the overall distribution of visual characteristics and patterns present in the image. In this context, the term 'visual vocabulary' simply means centroids generated by k-means clustering.

$$\mathrm{dist}(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4}$$

where, $P_1(x_1, y_1)$ represents the first point, $P_2(x_2, y_2)$ represents the second point.

$$V_i = \frac{1}{c_1} \sum_{j=1}^{c_1} X_j \tag{5}$$

where, V_i represents the centroid of cluster i, c_i represents the count or number of data points in cluster i, X_j represents the j-th data point assigned to cluster i.

In the final step of the training phase, a linear SVM was utilized to fit the generated visual words to the corresponding identity document classes. The linear SVM algorithm aimed to learn a decision boundary that effectively separated the different classes based on the visual word representations. By training the SVM on the labeled identity document samples and their associated visual word histograms, the classifier learned to distinguish between different document classes. This trained SVM model could then be used to predict the class of identity documents based on their visual word histograms.

In the testing phase of our experiment, we implemented three different strategies to evaluate the performance of our models. The first strategy, known as same-variant testing, involved utilizing each model to predict samples from the training dataset that shared the same variant as the dataset it was trained on. For instance, SGM-A was used to predict samples from SGM-B. The second strategy, cross-variant testing, required each model to predict samples from a training dataset with a different variant. For example, SGM-A was used to predict samples from USGM-B. Additionally, we performed k-fold cross-validation using the merged dataset, where k was set to 10. This involves dividing the dataset into 10 subsets or folds. In each iteration of the process, one fold was reserved for testing, while the remaining nine folds were used to train the model.

3.3 Testing phase

The testing phase consists of several key steps, including pre-processing, feature detection, vector quantization, and prediction. During the pre-processing step, images or samples with low visibility scores were excluded, and the test samples were resized to a proportional width of 320 pixels. Feature detection was performed using the SIFT algorithm. To simplify the representation of the test sample and improve efficiency, we applied vector quantization on the detected descriptors. Vector quantization assigns the test sample's descriptors to the nearest visual words in the visual vocabulary obtained during training. This step transforms the test sample into a collection of quantized descriptors, enabling compatibility with the trained model. Finally, the previously trained model, which had learned to classify the visual word representations of identity documents, was utilized to predict the class of the test sample based on its quantized descriptors. This model had been trained on labeled data during the training phase, allowing it to classify the test sample into one of the pre-defined classes.

To evaluate the performance of our models, we used accuracy as the metric. Accuracy, defined in Eq. (6), was calculated by dividing the number of correct predictions by the total number of processed samples. It is important to note that in our evaluation, true negatives and false negatives were not considered, making accuracy in our case can be considered as precision.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{6}$$

4. EXPERIMENT RESULT & DISCUSSION

Our experiment was conducted on a laptop with the following hardware specifications: Lenovo Legion 5 15ARH05, featuring an AMD Ryzen 7 4800H 2.90 GHz 64-bit processor and 16.0 GB of RAM. For our software setup, we utilized the Windows 11 Home Single Language 64-bit operating system, Spyder IDE version 5.3.0, Python version 3.9.5, and the opencv-contrib-python-headless module version 4.5.5.64.

The results of the same-variant testing are presented in Table 3. Based on our observations, the model with a cluster size of 192 generally achieved the highest accuracy across all dataset variants. Specifically, the best accuracy was attained by the SGM-A model when tested on SGM-B, achieving an accuracy of 98.5%. Moreover, we observed that accuracy tended to increase as the cluster size was increased. This suggests that, by increasing the cluster size in k-means, we can improve the accuracy of the models.

The results of the cross-variant testing are displayed in Table 4. For this test, we exclusively used the model with a cluster size of 192. This decision was based on our findings from the same-variant testing, where the model with a cluster size of 192 consistently achieved the highest accuracy across all dataset variants.

From Table 4, we observed that the performance of the segmented models, namely SGM-A and SGM2-A, performed well when predicting samples from other segmented datasets. For instance, when SGM-A was used to predict SGM2-B, it achieved an accuracy of 98.1%. Similarly, SGM2-A, when utilized to predict SGM-B, achieved an accuracy of 94.5%. However, their performance significantly dropped when utilized to predict samples from the unsegmented variant, USGM-B. In contrast, the performance of the USGM-A model exhibited relatively stable results. It achieved an accuracy of 96.9% when predicting samples from SGM-B and 93.9% when predicting samples from SGM2-B. While USGM-A may not have achieved the highest accuracy in this experiment, it is important to note that its accuracy remained more consistent compared to the segmented models.

The results of the k-fold cross-validation are presented in Table 5. Among the different cluster sizes tested, datasets with a cluster size of 192 generally demonstrated better performance. It is worth noting that the SGM2 dataset showed the highest average accuracy. However, we consider it to be unreliable due to its method of generation. The SGM2 dataset was divided or split based on the clarity of the identity documents, resulting in a merged dataset that essentially represents SGM but with a reduced number of samples. Therefore, caution should be exercised when interpreting the results of the SGM2 dataset within this specific context.

Table 3. Results of same-variant testing

Model	**Testing Dataset**	**Accuracy by Cluster Size (*k*)**			
		96	**128**	**160**	**192**
SGM-A	SGM-B	0.970	0.973	0.980	0.985
SGM2-A	SGM2-B	0.811	0.825	0.852	0.865
USGM-A	USGM-B	0.940	0.955	0.962	0.963

Table 4. Results of cross-variant testing

Model	**Testing Dataset**		
	SGM-B	**SGM2-B**	**USGM-B**
SGM-A	n/a	0.981	0.667
SGM2-A	0.945	n/a	0.705
USGM-A	0.969	0.939	n/a

Table 5. Results of k-fold cross-validation

Dataset	**Accuracy by Cluster Size (*k*)**			
	96	**128**	**160**	**192**
SGM	0.969	0.975	0.976	0.979
SGM2	0.988	0.990	0.992	0.993
USGM	0.949	0.961	0.967	0.972

What is particularly interesting is when we compare the k-fold cross-validation results of SGM and USGM datasets in Table 5 to cross-variant results in Table 4, the accuracy of the

SGM dataset is significantly lower when attempting to predict unsegmented samples. On the other hand, the accuracy of the USGM dataset remains relatively stable across both cross-variant testing and k-fold cross-validation.

Table 6. Method comparison

Model	Method	Testing Dataset		
		SGM-B	SGM2-B	USGM-B
SGM-A	SURF + k-means + SVM	0.976	0.977	0.369
	Proposed Method	**0.985**	**0.981**	**0.667**
SGM2-A	SURF + k-means + SVM	**0.964**	**0.913**	0.526
	Proposed Method	0.945	0.865	**0.705**
USGM-A	SURF + k-means + SVM	0.892	0.878	**0.979**
	Proposed Method	**0.969**	**0.939**	0.963

In Table 6, we evaluate our proposed method by comparing it to other classification methods. c. In both methods, k-means cluster size 192 was used. We observed that models trained on segmented datasets (SGM-A and SGM2-A) achieved higher accuracy when predicting other segmented samples but showed significantly lower accuracy when applied to unsegmented datasets, like USGM-B. In contrast, models trained on the unsegmented dataset, USGM-A, demonstrated more consistent accuracy.

Additionally, our proposed BoVW method, utilizing SIFT instead of SURF, achieved higher accuracy in 5 out of 9 test cases. This corresponds to SIFT's ability to extract a greater number of feature descriptors and being more resistant to noise and lighting variations present in the samples. The generation of fewer feature descriptors might result in a fewer distinctive visual vocabulary generated during the k-means process.

While SIFT requires more computational resources for both extraction and processing compared to SURF, we argue that higher accuracy can justify this trade-off. Failing to accurately determine an identity document can result in delays in the company's KYC process, leading to frustration for both users and service providers. Moreover, we expect the service provider to deploy the algorithm within a server environment rather than integrating it into the client-side process. This deployment strategy aims to leverage the greater computational resources available on servers for effectively managing the classification task. Such deployment allows the model to dynamically expand and scale as necessary, utilizing the computational capacity provided by server environments.

From this experiment, we have observed several key findings:

(1) We have demonstrated that high accuracy can be achieved in identity document classification tasks without the need for geometry transformations. This finding is consistent with the usage of SIFT in image matching and stitching, where SIFT is known for detecting features regardless of the orientation and distortion present in the image.

(2) Segmentation is encouraged only on the testing dataset or sample for which you want to predict its class, as it helps in trimming out irrelevant parts to be matched. However, it is not recommended to include segmentation in the training dataset because unsegmented samples may assist the algorithm in identifying and learning to ignore irrelevant parts. This conclusion is supported by the results obtained in the cross-variant testing (see Table 4), where the USGM-A model consistently achieves high accuracy when used to predict samples in SGM-B and SGM2-B.

(3) When comparing the results of SGM-A and SGM2-A in same-variant testing (see Table 3), it shows that SGM-A generally achieves better accuracy compared to SGM2-A in all cluster sizes. This is likely because the SGM-A covers a wider range of identity document conditions. This suggests that adding more variation in how an identity document was taken, could improve the model's ability to predict class.

(4) In both same-variant testing (see Table 3) and k-fold cross-validation (see Table 5), we observe a trend of increasing accuracy as the cluster size in k-means increases. The difference in accuracy is particularly significant from cluster size 96 to 160, but the impact becomes less pronounced from cluster size 160 to 192. This suggests that increasing the cluster size of k-means can potentially improve the accuracy of the model. However, it is important to consider the trade-off, as larger cluster sizes require additional time and computational resources to build the model.

(5) In k-fold cross-validation (see Table 5), both SGM and USGM achieve high average accuracy. However, in cross-variant testing (see Table 4), the accuracy of SGM-A significantly drops when used to predict samples from USGM-B. On the other hand, the accuracy of USGM-A remains relatively the same when used to predict samples from SGM-B and SGM2-B. This suggests that models trained using segmented datasets might be overfitting.

5. CONCLUSION AND SUGGESTIONS FOR FUTURE WORKS

In this paper, we have experimented with a classification approach for identity document images. The classification method involves categorizing identity documents based on their document type and the issuing country. However, one of the major challenges encountered in this task is the limited availability of discriminant features for accurate classification. Despite this challenge, the proposed method has demonstrated relatively stable accuracy on one of the dataset variants, achieving up to 97.2%.

Through our observations, we have noticed that introducing more variations in the training dataset, such as distorted, disoriented, and low-light identity documents, can enhance the performance of our model. It is important to acknowledge that the dataset used in our experiments has certain limitations. For example, every sample in a class within our dataset originates from a single identity document containing a fixed set of information. This scenario might lead to the model learning textual information as part of the image features.

To address this, future work should focus on generating variations within each class using pre-fabricated information. What this means is that research can be conducted to generate more variations of identity documents using fabricated information of an individual. This would not only increase variation in textual information but also enhance reproducibility while avoiding potential legal concerns.

Another finding indicates that by increasing cluster size in k-means we could improve model accuracy. However, it should be noted that more time and computational resources are required to build the model. Furthermore, the difference between the performance of the two upper clusters is not very significant. Additionally, we acknowledge the limitations of our research regarding the selection of cluster sizes. Which were based on a maximum width limit previously established. Therefore, further exploration of methods to optimize descriptor clustering into visual words, such as the elbow method, silhouette coefficient, and others, can be pursued.

We also discovered that refraining from performing segmentation on the training data benefits the algorithm by facilitating its learning process regarding which parts of the image are irrelevant. On the contrary, we encourage segmentation on the test data to eliminate irrelevant portions that may influence the classification of the test samples. Consequently, better methods for detecting and segmenting identity documents are still needed to enhance the overall classification process.

ACKNOWLEDGMENT

This research was conducted as the first author's thesis at Bina Nusantara University. I am grateful for the insights and guidance provided by Mr. Iman Herwidiana Kartowisastro, who served as the second author of this research paper. His expertise and support were instrumental in the successful completion of this work.

REFERENCES

[1] Castelblanco, A., Solano, J., Lopez, C., Rivera, E., Tengana, L., Martin, O. (2020). Machine learning techniques for identity document verification in uncontrolled environments: A case study. In Pattern Recognition: 12th Mexican Conference, MCPR 2020, Morelia, Mexico, June 24–27, 2020, Proceedings 12, pp. 271-281. https://doi.org/10.1007/978-3-030-49076-8_26

[2] Bulatov, K., Matalov, D., Arlazarov, V.V. (2020). MIDV-2019: Challenges of the modern mobile-based document OCR. In Twelfth International Conference on Machine Vision (ICMV 2019), pp. 717-722. https://doi.org/10.1117/12.2558438

[3] Rajput, V.U. (2013). Research on know your customer (KYC). International Journal of Scientific and Research Publications, 3(7): 541-546.

[4] Nguyen, T. T. T., Le Hong, L., Nguyen, H. N. (2020). An efficient method for automatic recognizing text fields on identification card. VNU Journal of Science: Mathematics-Physics, 36(1). https://doi.org/10.25073/2588-1124/vnumap.4456

[5] De Las Heras, L.P., Terrades, O.R., Llados, J., Fernandez-Mota, D., Canero, C. (2015). Use case visual bag-of-words techniques for camera based identity document classification. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, pp. 721-725. https://doi.org/10.1109/ICDAR.2015.7333856

[6] Awal, A.M., Ghanmi, N., Sicre, R., Furon, T. (2017). Complex document classification and localization application on identity document images. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, pp. 426-431. https://doi.org/10.1109/ICDAR.2017.77

[7] Yakovleva, O., Nikolaieva, K. (2020). Research of descriptor based image normalization and comparative analysis of SURF, SIFT, BRISK, ORB, KAZE, AKAZE descriptors. Advanced Information Systems, 4(4): 89-101. https://doi.org/10.20998/2522-9052.2020.4.13

[8] Chater, A., Lasfar, A. (2020). New approach to the identification of the easy expression recognition system by robust techniques (SIFT, PCA-SIFT, ASIFT and SURF). Telkomnika (Telecommunication Computing Electronics and Control), 18(2): 695-704. http://doi.org/10.12928/telkomnika.v18i2.13726

[9] Simon, M., Rodner, E., Denzler, J. (2015). Fine-grained classification of identity document types with only one example. In 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, pp. 126-129. https://doi.org/10.1109/MVA.2015.7153149

[10] Khandan, N. (2021). An intelligent hybrid model for identity document classification. arXiv preprint arXiv:2106.04345. https://doi.org/10.48550/arXiv.2106.04345

[11] Sicre, R., Awal, A.M., Furon, T. (2017). Identity documents classification as an image classification problem. In Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19, pp. 602-613. https://doi.org/10.1007/978-3-319-68548-9_55

[12] Arlazarov, V.V., Bulatov, K.B., Chernov, T.S., Arlazarov, V.L. (2019). MIDV-500: A dataset for identity document analysis and recognition on mobile devices in video stream. Computer Vision and Pattern Recognition, 43(5): 818-824. https://doi.org/10.48550/arXiv.1807.05786

[13] Karim, A.A.A., Sameer, R.A. (2018). Image classification using Bag of Visual Words (BOVW). Al-Nahrain Journal of Science, 21(4): 76-82. https://doi.org/10.22401/ANJS.21.4.11

[14] Mittal, H., Saraswat, M. (2019). Classification of histopathological images through bag-of-visual-words and gravitational search algorithm. In: Bansal, J., Das, K., Nagar, A., Deep, K., Ojha, A. (eds) Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, Springer, Singapore, 817: 231-241. https://doi.org/10.1007/978-981-13-1595-4_18

[15] Gao, Y., Lee, H.J. (2018). Car manufacturer and model recognition based on Scale Invariant Feature Transform. International Journal of Computational Vision and Robotics, 8(1): 32-41. https://doi.org/10.1504/IJCVR.2018.090014

[16] Sarwar, A., Mehmood, Z., Saba, T., Qazi, K.A., Adnan, A., Jamal, H. (2019). A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a Support Vector Machine. Journal of Information Science, 45(1): 117-135. https://doi.org/10.1177/0165551518782825

[17] Chaganti, S.Y., Nanda, I., Pandi, K.R., Prudhvith, T.G., Kumar, N. (2020). Image classification using SVM and CNN. In 2020 International conference on computer science, engineering and applications (ICCSEA), Gunupur, India, pp. 1-5. https://doi.org/10.1109/ICCSEA49143.2020.9132851

[18] Chow, B.H.Y., Reyes-Aldasoro, C.C. (2021). Automatic gemstone classification using computer vision. Minerals, 12(1): 60. https://doi.org/10.3390/min12010060

[19] Tensmeyer, C., Martinez, T. (2017). Analysis of convolutional neural networks for document image classification. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR), Kyoto, Japan, pp. 388-393. https://doi.org/10.1109/ICDAR.2017.71

[20] Zhao, R., Mao, K. (2017). Fuzzy bag-of-words model for document representation. IEEE Transactions on Fuzzy Systems, 26(2): 794-804. https://doi.org/10.1109/TFUZZ.2017.2690222

[21] Yao, L., Mao, C., Luo, Y. (2019). Graph convolutional networks for text classification. In Proceedings of the AAAI conference on artificial intelligence, 33(1): 7370-7377. https://doi.org/10.1609/aaai.v33i01.3301737

[22] Asim, M.N., Khan, M.U.G., Malik, M.I., Razzaque, K., Dengel, A., Ahmed, S. (2019). Two stream deep network for document image classification. In 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, pp. 1410-1416. https://doi.org/10.1109/ICDAR.2019.00227

[23] Ghanmi, N., Awal, A.M. (2018). A new descriptor for pattern matching: application to identity document verification. In 2018 13th IAPR international workshop on document analysis systems (DAS), Vienna, Austria, pp. 375-380. https://doi.org/10.1109/DAS.2018.74