Improving classification of an industrial document image database by combining visual and textual features

Olivier Augereau, Nicholas Journet, Anne Vialard, Jean-Philippe Domenger

Improving classification of an industrial document image database by combining visual and textual features

Olivier Augereau
Gestform
38 Rue François Arago
33700 Merignac, France
Email: oaugereau@gestform.com

Nicholas Journet, Anne Vialard and Jean-Philippe Domenger
LaBRI - Laboratoire Bordelais de Recherche en Informatique
Univ. Bordeaux, UMR 5800
F-33400 Talence, France
Email: {journet, vialard, domenger}@labri.fr

***Abstract*—The main contribution of this paper is a new method for classifying document images by combining textual features extracted with the Bag of Words (BoW) technique and visual features extracted with the Bag of Visual Words (BoVW) technique. The BoVW is widely used within the computer vision community for scene classification or object recognition but few applications for the classification of entire document images have been submitted. While previous attempts have been showing disappointing results by combining visual and textual features with the Borda-count technique, we're proposing here a combination through *learning approach*. Experiments conducted on a 1925 document image industrial database reveal that this fusion scheme significantly improves the classification performances. Our concluding contribution deals with the choosing and tuning of the BoW and/or BoVW techniques in an industrial context.**

I. Introduction

The work presented in this paper takes place in an industrial context. A digitizing company such as Gestform [1] digitizes several millions of documents each month. They include many documents of different natures such as human resources documents (identity papers, payrolls, forms, etc.) or expense account (transport tickets, restaurant receipts, hotel bills, etc). These document images pose the following problems: some documents are damaged (cut, crumpled, torn, stained, etc..), printed on thin paper with ink that easily fades (such as restaurant tickets) or contain very few words. Figure 1 presents several examples of document images extracted from an industrial digitization process. For example, digitized bills and tickets ("Flunch, "Quick" and "Buffalo") are printed with low-quality papers and inks. Others images ("RATP" class) are very small, bent or folded subway tickets.

The document image quality is not the only complication for a classification purpose. One other issue is to gather in the same class documents with some visual differences (layout, text, image). For example the classes "Buffalo", "Flunch" and "Quick" are composed of bills that have different sizes or text content. In the same way, the class "IBIS" contains hotel bills where most of the elements show variations due to their different hotels of origin (logo, text, layout...). Another complex scenario appears when document images with high visual similarities need to be classified in different classes. For

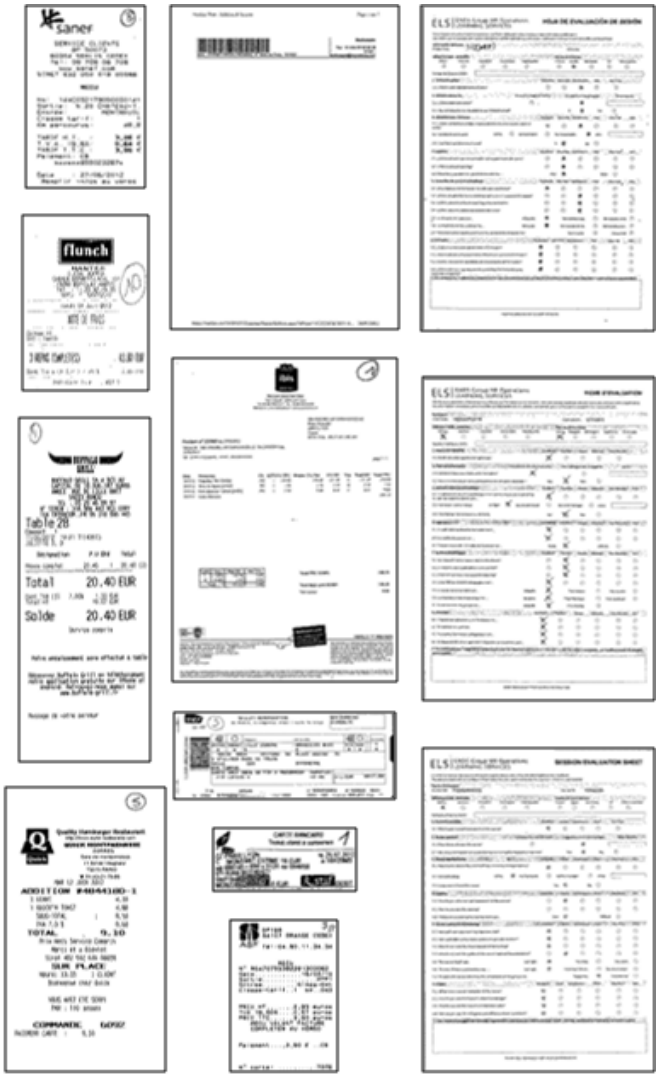

Fig. 1. Examples of 12 different classes of documents. From top to bottom and left to right: "ElsUk", "ElsEs", "ElsFr", "IBIS", "Lot", "ASF", "SANEF", "SNCF", "Flunch", "Quick", "Buffalo" and "RATP". Quality image have been reduced for confidentiality reasons. These documents will be used for the tests in the last part of this paper.

[1] www.gestform.com

example, the three classes "ElsUk", "ElsEs" and "ElsFr" are different types of check-box forms that look very similar. The difference is mainly the language as the forms are respectively written in English, Spanish and French.

Fig. 2. Complexity of text and layout extraction: two original images from "RATP" class (row 1), the extracted structures (row 2) and the OCR results (raw 3). The layout and the OCR are carried out by FineReader 10. For these two very similar documents, the layout and text extracted are very different.

Figure 2 illustrates the limitations of document images classification based on text or layout analysis. In this example the layout and text extracted from very similar documents are different. In this context, where some documents are degraded and some show a layout which is difficult to extract, good OCR results are hard to obtain. This observation leads us to the conclusion that using image features could improve textual based classification.

In this paper we propose to apply the BoVW technique for the classification of entire document images classification and to combine it to the BoW results in order to improve the performances for document images classification. We point out the importance of choosing a pertinent fusion technique. Another contribution of this paper is a detailed discussion on how to apply this classification scheme in an industrial context.

There will firstly be a presentation of related works, then the BoW and BoVW techniques will be described as well as a way to combine them. Finally, the results and the usage of the features in an industrial context will be discussed.

II. Related works

In the survey [1], Chen and Blostein present several techniques of classification of document images. Three main groups of features used for the classification task are distinguished: the text (words, numbers, etc.) extracted by an OCR, the image (colors, textures, shapes, etc.) and the layout (description of the document images structure). It can also be pointed out that most of the techniques presented in the survey are supervised. We also think that the supervised learning can help overcome the complexity of the database because it allows handling the diversity of the documents within a class and the dissimilarity between classes. The BoW and BoVW techniques are both supervised techniques.

A. The bag of words

The Bow technique is a well known method used for text classification. It consists in resuming a text as a vector measuring the frequency of a set of words. In 2002, two major works about machine learning for text categorization - the survey of Sebastiani [2] and the book of Joachims [3] - show that the bag of words approach is one of the best existing method for automatic text classification.

BoW are used for many tasks based of text analysis such as recommender systems, email classification, sentiment analysis and opinion mining [4], etc.

B. The bag of visual words

The BoVW method is more and more used in the computer vision community for natural images classification [5], object recognition [6] or CBIR (Content Based Image Retrieval) [7]. The successful applications on natural images leads other communities to use it. However, few document image applications using BoVW technique have been proposed. The BoVW technique has been applied to logo recognition [8], word spotting [9] and handwritten character recognition [10].

We can notice that, in document image field, the BoVW technique has mainly been applied to document images subparts retrieval. As far as we know, very few applications using the BoVW technique for classifying "whole" document images have been proposed. Recently, the authors of [11] detailed a multipage document retrieval system based on textual and/or visual features. They conclude that the visual features lower the retrieval performances when they are combined with textual features.

C. Multimodal fusion techniques

Many studies have already tackled the subject of multimodal data processing. For example, in the field of video analysis, three distinct modes are generally distinguished: sound, image and text [12]. In order to automatically combine the different modes, two main approaches stand out: the early fusion and the late fusion. The early fusion approach consists in finding a way to combine the features of the modes before the classification. For example, if the features are a vector, a simple way to do early fusion is to concatenate the features. In the late fusion approach, the features of each mode are learned separately and then their output are combined. A simple way to do late fusion for classification is to compute a vote of each classifier.

According to Meüller [13], late fusion performs better than early fusion in many experiments applied in multimodal information retrieval, based on the combination of visual and textual features. The three major schemes of late fusion are:

1) Combining scores or ranks [14] (such as Borda-count). The classifiers output a rank for each class, then the combination is then done by applying a re-ranking.
2) Combining class probabilities with rules (average, sum, product, etc.) [15]. The classifiers output a probability for each class, the combination is done by summing or multiplying the probabilities.

3) Combining by learning [16]. Each classifier outputs a measure for each class (probability, distance, etc). Then, another learning is carried out by using all these measures as a feature vector.

The author of [16] tested 38 combination schemes of SVM for handwritten digit recognition. It clearly appears that the combination by learning outperforms the rank combination and the rule combination.

III. Our implementation of document classification

A. Classification with BoW

The BoW technique relies on the principle that a document can be described by counting the occurrences of a defined set of words.

In a first step, OCR is applied on document images in order to extract words. Some preprocessing are applied to the text. Special characters such as punctuation are removed. Words which contain less than 4 letters or more than 15 are removed. This step is a little rough, many words such as articles, some conjunctions, some pronouns and isolated noise are filtered , for examples : "the, a, de, le, la, les", "or, but, and, ou, et, or, ni, car", "she, he, my, je, tu, il" and "tfx, zxm, uu". We also observed that, most of the time, words with more than 15 letters are due to OCR errors. After this, three other steps are usually applied: stop word removal, stemming and/or lemmatization. The stop words are the most common words, their frequency in each document is considered as not discriminating. Stemming and lemmatization consist in replacing the different inflected forms of a word into a single term. Our results do not need this steps because we are dealing with tickets, receipts, etc. which contain very few sentences. After the preprocessing, a dictionary is created in order to define the set of words that is relevant to count. The 1000 most frequent words are selected for building this dictionary. Uncommon OCR errors are not handled with this methods, but OCR tends to provide similar result with similar document images. So, if a word come up frequently with an error, most of the time he will have the same error and the word will be selected in the dictionary. The features selected to describe a document are the number of occurrences of each word of the dictionary. At last, each document is described by a histogram with 1000 bins. A SVM classifier is trained with the learning documents and then tested documents are classified.

B. Classification with BoVW

The BoVW technique relies on interest point detection such as SURF [17] or SIFT [6] which implies robustness to transformations such as rotation, translation, scaling, blur and illumination change.

BoVW models have been directly inspired by BoW technique. In most of applications, visual words are interest points. BoVW algorithm is described in figure 3. It can be sumed up in four main steps. First, interest points are extracted from each image to classify. Then, all these interest points are clustered in k clusters. In a third step, the image is described by a histogram of k bins. Each interest point of the image increments the bin of the histogram corresponding to its cluster. Finally, the images are classified by using a machine learning algorithm.

Usually, interest points are extracted with SIFT or SURF, the features clustering is done with the k-means algorithm and supervised learning is done with SVM [18], [7]. We used the SURF and k-means implementation of openCV library. Interest points are extracted and described on resized images (30% of original size). We chose $k = 1000$ for k-means. We use a multiclass SVM with RBF kernel. Parameters are auto-tuned by cross validation using libSVM library.

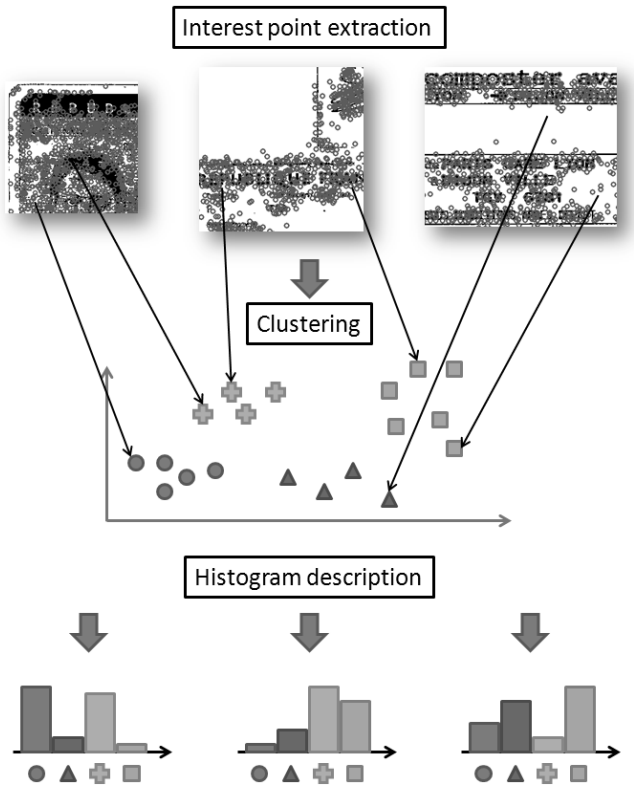

Fig. 3. Description of histograms of BoVW method. 1) Interest points are extracted on each image. 2) Interest points are clustered in k clusters . 3) Because each interest point is linked to one of the k clusters, each image is described by an histogram including the number of occurrences of each k classes.

C. Classification with fusion of BoW and BoVW

Figure 4 summarizes the late fusion technique that we propose. The BoW and the BoVW are applied separately. By default SVM don't return probabilities. That's why we use the Platt's method [19] which is one of the the most famous methods for converting the SVM outputs into probabilities. Then, the probabilities are concatenated, normalized and provided as an input of a new learning stage (another SVM classifier is used). The same learning images are used for the BoW, the BoVW and the fusion classifiers.

IV. Tests and results

The database is composed of 1985 documents digitized by Gestform company, randomly selected from a production chain. Images are digitized at 300 dpi and are automatically binarized by scanners. The database contains 12 classes of

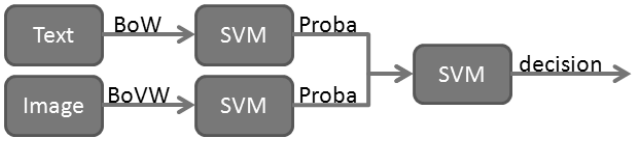

Fig. 4. Late fusion scheme. SVM are applied separately on BoW and BoVW. The probabilities output of SVM are concatenate and provided as the input of another SVM.

documents. One example of each class is shown in figure 1. Some classes of the database are challenging, details are provided in the introduction.

A recent work [11] concluded that the fusion of BoW and BoVW for a document image retrieval purpose is not pertinent. Globally, including image features reduces the results obtained by using only textual features. In our classification context, the same conclusion is obtained by the following experiment. We apply a late fusion based on combining ranks (Borda technique), the visual features decreases the textual feature results. Table I shows that the performances drop of 8% when we use textual and image features on our database instead of using only textual features.

TABLE I. AVERAGE RECALL AND PRECISION FOR BOW, BOVW AND BORDA FUSION TECHNIQUES.

	Recall	Precision
BoVW	0.871	0.869
BoW	0.982	0.977
Borda fusion	0.899	0.902

In the next subsection, the document image classifications using BoW, BoVW and fusion by learning are described. In table II, the results of classification are detailed class by class to highlight the cases where the visual features lead to a better classification than the textual features. Finally, the tests show that fusion by learning, instead of Borda fusion, can produce a pertinent signature for each document by using both visual and textual information.

A. Test protocol

Performances are computed using a cross-validation model validation technique. For each class, 5 documents are randomly selected for learning. All other 1925 documents are tested. This method has been applied 10 times, so the results (recall and precision) are an average of the 10 tests.

Usually, supervised methods use around 60% of each class for learning and 40% for testing. It can be noted that only 5 images are used for learning each class, here. There are two main reasons linked to the industrial context. The first one is that we don't know the database in advance, so we don't know how many documents compose each class. The second one is that manually labeling the documents is very time consuming. To be concretely applied, this step has to be reduced as much as possible.

B. Results

The results are displayed in table II. The number of the test documents per class is displayed in the second column (without the 5 documents per class used for learning).

The results show that BoW has very good performances. However, it misses some documents such has "IBIS" (hotel invoice) and "RATP" (subway tickets). For the "IBIS" class, the explanation is that the documents come from different hotels and have different layouts and texts. Furthermore, the text in the logo and the small print about general conditions cannot be read by an OCR. An OCR produces a low quality transcription on this kind of documents because they contain little text and have many degradations. For the "RATP" class, image quality is the main reason for a low recall value (figure 2).

Globally, the results of BoVW are good but not as much as the performances obtained by using BoW. The three classes "ElsEs", "ElsFr" and "ElsUk" are very confusing visually and thus highly degrade the final performances of BoVW. But, on some classes where BoW does not perform well (such as "IBIS" or "RATP"), BoVW works better.

The combination leads to two main results: 1) improvement of either BoW or BoVW or 2) improvement of both BoW and BoVW. If one technique (BoW or BoVW) works better than the other one, the combination may provide intermediate results (case 1). For example, for the class "IBIS", the precision of the fusion is better than the precision of BoW but is worse than the precision of BoVW. The explanation is that for a few documents, one classifier gives the good answer with a low confidence rate where the other classifier gives a wrong answer with a higher confidence rate.

In most cases, the fusion technique provides better or equal results than using either BoW or BoVW techniques. The "SNCF" class illustrates an interesting case where the fusion outperforms both BoW and BoVW precisions. This can be explained by the fact that, for a given class, the two classifiers provide right answers on two separate subparts of the class. Finally, it is also important to keep in mind that the gain of the last percentage in recall and precision (improving from 98% to 99% or from 99.90% to 99.99%) is the most difficult to get. In any case, there always will be ambiguous documents, so obtaining 100% of recall and precision is idealistic in an industrial context.

TABLE II. RECALL AND PRECISION FOR BOW, BOVW AND FUSION BY LEARNING. THE AVERAGES ARE WEIGHTED BY THE NUMBER OF DOCUMENTS.

		Recall			Precision		
Classes	Nb docs	BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0.984	0.995	0.964	0.954	0.990
ElsEs	299	1	0.677	1	0.987	0.709	0.997
ElsFr	99	1	0.816	1	1	0.626	1
ElsUk	430	1	0.782	1	1	0.802	1
IBIS	41	0.810	1	0.972	0.829	0.902	0.854
Lot	670	1	1	1	0.975	0.987	0.991
ASF	31	1	0.792	1	0.742	0.613	0.645
SANEF	22	1	0.167	1	0.955	0.682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0.667	0.231	0.400	1	0.750	1
Quick	12	1	0.923	1	1	1	1
RATP	110	0.764	0.940	0.940	1	1	1
Average		0.982	0.872	**0.994**	0.977	0.870	**0.986**

C. Feedback on an industrial implantation

Most of the time, applying research techniques in an industrial context is challenging. Many techniques are *ad hoc*

or based on rules. Indeed, it is difficult to apply them in many different contexts. Using statistics about textual and visual features allows to make classification working on many kinds of documents.

About parameters: BoW and BoVW do not need many parameters to work well and are quite simple to apply. We only have to choose three parameters: the number of clusters in k-means for BoVW, the size of the dictionary of BoW and the Hessian threshold for interest points extraction. The two first parameters have been coarsely fixed to 1000 and the Hessian threshold to 500. Changing the Hessian threshold has little impact on results because documents are binary. The number of clusters and the size of the dictionary have also little impact as we just need to have enough (visual) words in order to describe the documents and not too much in order not to be too specific. A range between 200 and 2000 seems to work well in most cases.

How to choose between BoW, BoVW and fusion: The tests carried-out in this paper show that for some classes it is better to use only BoW, for other ones to use only BoVW and for most of them to use the fusion. The choice between the three methods will be guided by the user knowledge, which can change depending on the context. If the documents show good quality and contain many words, the best performances will be obtained by using BoW. If the documents have defects, geometrical affine transformation, handwriting or any other characteristic that can make the OCR works poorly, the best performances will be obtained by using BoVW. Finally, if the database is totally unknown and it is difficult to predict whether the OCR will work well or not (which is common in an industrial context), we advise to use BoW and BoVW fusion.

Processing time: In average, applying OCR on documents takes around 3 seconds per image while extracting points (on 30% resized images) takes around 0.9 seconds. The learning stage is done offline so it does not matter if it is time consuming or not. The learning for BoW takes only a few minutes because it just consists in counting and filtering words. The learning step of BoVW takes a little more time because the k-means algorithm has to be applied on several millions interest points so a few hours are needed. Finally, the recognition is made by SVM which is extremely fast and takes around 20 milliseconds to take a decision.

V. Conclusion and perspectives

In this paper, we have shown that BoVW method, which is usually used for natural images classification, can also be applied to document images, providing interesting results when OCR fails to recover textual information.

Even if, generally, BoVW works a little less better than BoW, we show that using a combination (fusion by learning scheme) of both techniques improves recall and precision of document images classification.

Many perspectives remain for improving multimodal document images classification in an industrial context. For example, we would like to consider : testing others combination techniques; trying to use different learning databases for text and image features; choosing the best classifier for each document instead of systematically combining the classifiers; adding others features such as layout.

Acknowledgment

The authors would like to thank Jean-Marc Nahon, the computer science director of Gestform company and Ellina Guibert, a student at ENSEIRB-MATMECA graduate school of Bordeaux Institute of Technology. All images are provided by Gestform.

References

[1] N. Chen and D. Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *International Journal on Document Analysis and Recognition*, vol. 10, no. 1, pp. 1–16, 2007.

[2] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[3] T. Joachims, *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers Norwell, MA, USA:, 2002, vol. 186.

[4] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[5] J. Yang, Y. Jiang, A. Hauptmann, and C. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, 2007, pp. 197–206.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] E. Valle and M. Cord, "Advanced Techniques in CBIR: Local Descriptors, Visual Dictionaries and Bags of Features," in *Tutorials of the XXII Brazilian Symposium on Computer Graphics and Image Processing*. IEEE, 2009, pp. 72–78.

[8] M. Rusinol and J. Lladós, "Logo spotting by a bag-of-words approach for document categorization," in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 111–115.

[9] R. Shekhar and C. Jawahar, "Word image retrieval using bag of visual words," in *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*. IEEE, 2012, pp. 297–301.

[10] W. Song, S. Uchida, and M. Liwicki, "Look inside the world of parts of handwritten characters," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 784–788.

[11] M. Rusinol, D. Karatzas, A. D. Bagdanov, and J. Llados, "Multipage document retrieval by textual and visual representations," in *Proc. of International Conference on Pattern Recognition (ICPR)*, 2012. [Online]. Available: http://www.micc.unifi.it/publications/2012/RKBL12

[12] C. G. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia tools and applications*, vol. 25, no. 1, pp. 5–35, 2005.

[13] H. Meüller, P. Clough, T. Deselaers, and B. Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer, 2010, vol. 32.

[14] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 1, pp. 66–75, 1994.

[15] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[16] D. Gorgevik and D. Cakmakov, "Handwritten digit recognition by combining svm classifiers," in *Computer as a Tool, 2005. EUROCON 2005. The International Conference on*, vol. 2. IEEE, 2005, pp. 1393–1396.

[17] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[18] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, 2004, p. 22.

[19] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.