A Study of Real-Time Hand Gesture Recognition Using SIFT on Binary Images

Wei-Syun Lin[1], Yi-Leh Wu[1,*],Wei-Chih Hung[1], and Cheng-Yuan Tang[2]

[1] Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
ywu@csie.ntust.edu.tw
[2] Department of Information Management, Huafan University, Taiwan

Abstract. We present a novel way to use the Scale Invariance Feature Transform (SIFT) on binary images. As far as we know, we proposed employ SIFT on binary images for hand gesture recognition and provide more accurate result comparing to traditional template approaches. There exist many restrictions on template matching approaches, such as the rotation must be less than 15°, and the variation on scale, etc. However, our proposed approach is robust against rotations, scaling, illumination conditions, and can recognize hand gestures in real-time with only off-the-shelf camera such as webcams. The proposed approach employs the SIFT features on binary image, the k-means clustering to map keypoints into a unified dimensional histogram vector (bag-of-words), and the Support Vector Machine (SVM) to classify different hand gestures.

1 Introduction

Hand gesture plays an important role of a social communication bridge. Gestures are the motion of the body or physical action form to convey some meaningful information. The difficulties of hand gestures recognition are to recognize hand gestures in real time with the high degrees of freedom (DOF) of the human hand. The ideal hand gestures recognition system have to meet the requirements in terms of real-time performance, recognition accuracy, robustness against transformations, cluttered background, and with hands from different people.

This work presents a novel way to employ the Scale Invariance Feature Transform (SIFT) on binary images, to the best of our knowledge, this work is the first research to use the SIFT method on binary images for hand gesture recognition. The traditional template matching approaches face many restrictions on the design of templates, such as rotation, scaling, etc. The proposed approach is robust against rotation, scaling, illumination conditions, and with real-time performance for hand gesture recognition.

In [1], Dardas et al. proposed a new technique to detect hand gestures only using face subtraction, skin detection, and hand posture contour comparison algorithm. However, Dardas et al. focused on bare hand gesture recognition without the help of

[*] Corresponding author.

J.-S. Pan et al. (Eds.): *Advances in Intelligent Systems & Applications*, SIST 21, pp. 235–246.
DOI: 10.1007/978-3-642-35473-1_24

any markers and gloves but required additional trainings to recognize hands from different people [1]. Dardas et al. employed the Viola–Jones method [11], which is considered the fastest and most accurate learning based method, to detect faces in images [1]. The detected face will be subtracted by replacing the face area with a black circle. After subtracting the face, Dardas et al. detected the skin area using the hue, saturation, value (HSV) color model. The proposed method has real-time performance is robust against rotation, scaling, and lighting condition changes. Then, the contours of skin area were compared with all the predefined hand gesture contours to remove other skin-like objects in the image.

Given an input image, Lowe's method [12] extracts a large collection of feature vectors, each of which is invariant to image translation, scaling, and rotation, partially invariant to illumination changes and robust to local geometric distortion. Therefore, the SIFT is adopted in this work for the bare hand gesture recognition. However, the SIFT features are of too high dimensionality to be used efficiently. The work proposes to alleviate the high dimensionality problem by employing the bag-of-features approach [13-14] to reduce the dimensionality of the feature space.

Many gesture recognition techniques were developed for vision-based hand gesture recognition with different pros and cons. The traditional approaches are the template based hand pose recognition and the appearance based features of hand [2-3]. These approaches have real-time performance because the easier 2-D image features are employed. There are three main steps in the traditional hand gesture recognition approaches: the hand segmentation, the feature extraction, and the posture recognition. However, the traditional hand gesture recognition approaches are very sensitive to illumination conditions.

There are some other approaches for hand gesture recognition which employ additional sensors to collect data and the recognition performance increases with more details of the collected data. But generally speaking, the more details of the collected data increase the processing cost. In [5], Van den Bergh et al. proposed a hand gesture interaction system based on a RGB camera and a Time-of-Flight (ToF) camera for real-time hand gesture interaction with high recognition accuracy. Their proposed system may have many advantages; however, the cost of ToF cameras is extremely high when comparing web-cams, the ones employed in our proposed approach.

Xu et al. proposed a hand gesture recognition system which utilizes both the multi-channel surface electromyogram (EMG) sensors and the 3D accelerometer (ACC) to achieve the average recognition accuracy about 91.7% in real application [6]. However, the time delay between the finished gesture command and the system response as a cube action is about 300ms, which is slower than our proposed approach.

Our proposed method is mainly based on the method proposed in [1] with the main idea of using machine learning algorithms to train and test different hand gesture models. But when the training is imperfect, the machine learning algorithms tend to produce inferior recognition results. For example, if there is a hand gesture that does not exist in the training model or the illumination conditions differ between the training and the testing model, the recognition accuracy will be reduced.

The main overview of [1] is as shown in Fig. 1. The input image captured by the webcam is transformed to gray scale and directly extract the keypoints in the input image using SIFT. The vector quantization (VQ) maps the keypoints of every training image into a unified dimensional histogram vector after the K-means clustering. The histogram is then employed as the input data for the SVM classification. The main problem of this approach is that the SIFT algorithm is a robust algorithm which extracts many local features as keypoints. And too many training keypoints will leads the machine learning process to over-fit the learned models. So this work proposes a new approach using SIFT on images to alleviate the above problem.

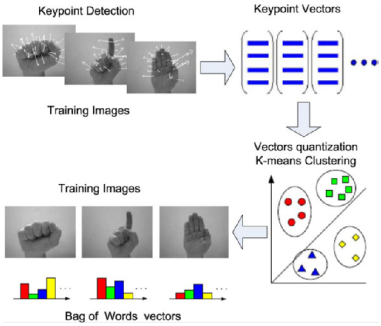

Fig. 1. Dardas el al. method [1] to generate the bag-of-words vectors

The major contributions of this work are as follows:

I. To our knowledge, this work is the first to employ the SIFT on binary images, which is considered infeasible in the past. The experiment results suggest that the proposed approach increases the robustness to recognize hands from different people.

II. The proposed approach is also robust against rotation, and scaling, unlike the traditional template matching approaches.

III. By removing unnecessary features and leaving only the useful features, the proposed system can achieve real-time performance for hand gesture recognition with high accuracy.

2 System Overview

Compared with related works, the main difference in the proposed work is to employ the SIFT on binary images instead of color/grey images. The proposed method removes the redundant information to achieve real-time and high recognition accuracy with different training hands (the hand features do not exist in the bag of features). Our hand gesture recognition system consists of two stages: the offline training stage and the online testing stage. All the images in the training stage and the testing stage are captured from a webcam.

2.1 Training Stage

In the training stage, the hand gesture training images can be represented by sets of keypoint descriptors. However, the numbers of keypoints from individual images are different and the keypoints lack meaningful order. The variant number of features and the lack of feature order create difficulties for machine learning methods such as the multiclass support vector machine (SVM) classifier that require feature vectors of fixed dimension as input. To address this problem, this work proposes to employ the bag-of-features approach, which has several steps.

The first steps is to extract the features (keypoints) from hand gesture training images using the SIFT algorithm. But the SIFT is sensitive to local geometric distortion, so the SIFT will generate different keypoints from images with different people. The features of hand shape and the features of between fingers are important for hand gesture recognition. But the other hand features such as the fingernail, fingerprint, and the skin color that differ from person to person are not important for hand gesture recognition. For this reason, this work proposes a new way that employs the SIFT on binary images.

In our approach, the first step is to transform the input hand gesture images to binary images. The SIFT method is then employed on the binarized hand gesture images to extract only the important hand features for gesture recognition. The next step is to employ the vector quantization (VQ) technique [2], which clusters the keypoint descriptors in their feature space into a number of clusters using the K-means clustering algorithm. Then each keypoint in encoded by the index of the cluster (codevector) to which this keypoint belongs. This VQ process maps keypoints of every training image into a unified dimensional histogram vector as shown in Fig. 2. Finally, each cluster is considered as a visual word (codevector) that stands for a particular local pattern shared by the keypoints in that cluster.

The clustering process constructs a visual word vocabulary (codebook) representing the local patterns in the training images. The size of the vocabulary is

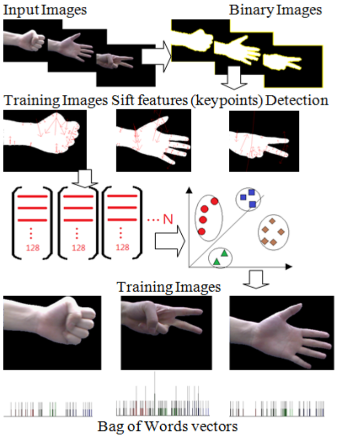

Fig. 2. Generating the bag-of-words vectors for training

determined by the number of clusters (codebook size), which can be varied from hundred to over tens of thousands. Each training image can be described as a "bag-of-words" vector by mapping the many keypoints to one visual word vector. The multiclass SVM models can be trained with the unified dimensional feature representation of visual word vectors. This work assumes that by employing the Dardas et al. method [1] to detect and track the hands and the skin color in images, the input image can be separated into foreground and background. And the training images contain only the hand gestures without any other objects, such as elbow or torso, in the input image. The size of the training images is by default 640×480 pixels in size but any other image sizes will work.

2.1.1 Binary Images

The work assumes that there is only one hand gesture in an input image and the hand gesture image is already divided into foreground (hand) and background (other) by applying method such as [1]. In this case, the foreground (hand gesture) of the input image is set to white and the background (other) is black.

With the above assumption, we effectively constraint some variant conditions such as illumination and skin colors of different test subjects. Thus the resulting SIFT features are inherently more robust. Fig. 3 shows the differences between applying the SIFT on binary images and applying the SIFT on color/grey images. Notice the redundant features vectors on fingernail, fingerprint, knuckle, etc., when applying the SIFT method on color/grey images as shown in Fig. 3(a) and 3(c). So using SIFT on binary image is useful, it can remove most redundant vector. When applying the SIFT method on the binary image as shown in Fig. 3(b) and (d), notice that only the most discriminating SIFT features, between fingers or the shape of hand, are remained. These most discriminating SIFT features are expected to increase the robustness of the proposed hand gesture recognition method.

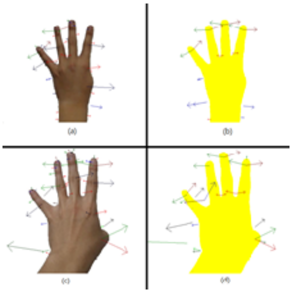

Fig. 3. SIFT extract from hand (a) SIFT keypoints from color image (b) SIFT keypoints from binary image (c) SIFT keypoints from color image (d) SIFT keypoints from binary image

2.1.2 SIFT

Lowe proposed the Scale- Invariant Feature Transform (SIFT) in [12]. The SIFT descriptors describe the local feature in the image geometric variations.

2.1.3 K-means Clustering

In data mining, the k-means clustering is a method of cluster analysis which partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The k-means clustering is an approach of unsupervised learning algorithm and an ordinary method for statistical data analysis applied in several fields. In the training stage, when the training images contain only hand gestures in color on a white background, the extracted keypoints represent the hand gesture only. The number of extracted keypoints normally does not exceed 75 for each gesture. But there are some redundant keypoints that have to be removed from the color hand gestures. When applying the SIFT on binary hand gesture images, the maximum number of extracted keypoints decreased to 50.

When using the SIFT keypoints extracted from binary images, if the number of clusters is too small, the classification accuracy will decrease. In the training image set, each gesture has 100 training images and the total number of keypoints for each gesture is about 5000. Because the each keypoints in the training image set is important and unique, we chose the value 800 as the number of clusters (visual vocabularies or codebook) to build our cluster model.

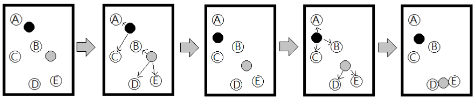

Fig. 4. K-means clustering with two clusters and saving the cluster model

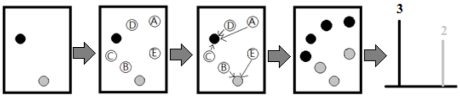

Fig. 5. Using cluster model to maps keypoints into a unified dimensional histogram

The first step in the k-means clustering is to divide the vector space (128-dimensional feature vector) into ***k*** clusters. The k-means clustering starts with ***k*** randomly located centroids (points in space that represent the center of the cluster) and assigns every keypoint to the nearest cluster centroids. After the assignment, the centroids (codevectors) are shifted to the average location of all the keypoints assigned to the same cluster, and the assignments are redone. This procedure repeats until the assignments stop changing. Fig. 4 shows this process in action for five keypoints: A, B, C, D, and E to form two clusters.

Fig. 5 shows the process that maps the image keypoints to a ***k*** dimensional histogram vector. By using the above k-means clustering method, the keypoint vectors for each training image are employed to build the cluster mode. The number of clusters (codebook) will represent the number of centroids in the cluster model. Finally, the cluster model will build codevectors equal to the number of clusters assigned (***k***) and each codevector will have 128 components, which is equal to the length of each keypoint. Then, the keypoints of each training image are mapped into the k-means clustering model to reduce the dimensionality into one bag-of-words vector with ***k*** components, where ***k*** is the number of clusters. In this way, each keypoint, extracted from a training image, will be represented by one component in the generated bag-of-words vector with value equal to the index of the centroids in the cluster model with the nearest Euclidean distance. The generated bag-of-words vector, which represents the training image, will be grouped with all the generated vectors of other training images that have the same hand gesture and labeled with the same number, and this label will represent the hand gesture class number. For example, label or class 1 for the hand gesture **C** training images, class 2 for hand gesture **fist**, class 3 for hand gesture **five**, and class 4 for hand gesture **index**.

By analyzing the SIFT keypoints, we discover unique characteristics of the SIFT vectors for each gesture. In Fig. 6, the SIFT keypoints of every gestures have some common regular patterns. But the VQ process maps the keypoints of a training image into one *k* dimensional histogram vector after K-means clustering. And with the k-means clusters, similar vectors will be mapped into the same dimension as show in Fig. 7. And if there are some minor differences in the SIFT vectors in the hand gesture image taken from different people, the k-means clustering will maps the similar gesture vectors to the same cluster.

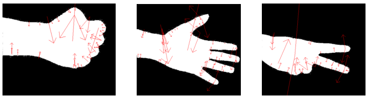

Fig. 6. SIFT vectors of every gesture

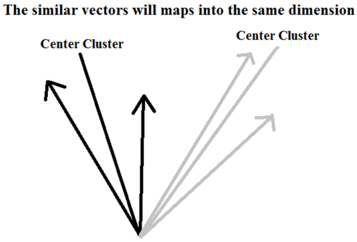

Fig. 7. Hand SIFT vectors mapped with k-mean

2.1.4 K-means Clustering

After mapping the keypoints of each training image to one bag-of-words vector, the bag-of-words vector is labeled with the hand gesture class or label number. All the labeled bag-of-words vectors are employed as the training data to build the multi-class SVM classifier model. The SVM is a supervised learning method for classification and regression by creating an n-dimensional hyperplane that optimally divides the data into difference groups. Even though SVMs were initially intended as binary classifiers, other methods that deal with a multiclass problem as a single "all-together" optimization problem exist [7], but are computationally much more costly than solving several binary problems.

A variety of approaches for decomposition of the multiclass problem into several binary problems using the two-class SVM have been proposed. In our implementation, multiclass SVM training and testing are performed using the LIBSVM library [8]. The LIBSVM supports multiclass classification and uses a one-against-one (OAO) approach for multiclass classification in SVM [9]. For the M-class problems (M being greater than 2), the OAO approach creates M(M-1)/2 two-class classifiers, using all the binary pair-wise combinations of the M classes. Each classifier is trained using the samples of the first class as positive examples and the samples of the second class as negative examples. To combine these classifiers, the Max Wins method is used to find the resultant class by selecting the class voted by the majority of the classifiers [10].

2.2 Testing Stage

The testing stage workflow is shown in Fig. 8. The first step is to capture image frames from the webcam or video file. And then apply the Dardas et al. method [1] for face detection and subtraction, hand gesture detection, and hand extraction in each image frame. These extracted hand images are employed for the testing model.

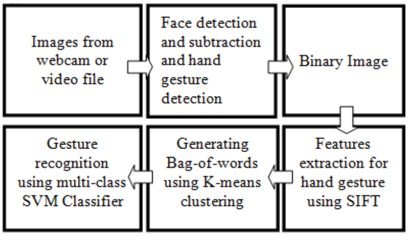
Images from webcam or video file
Face detection and subtraction and hand gesture detection
Binary Image
Features extraction for hand gesture using SIFT
Generating Bag-of-words using K-means clustering
Gesture recognition using multi-class SVM Classifier

Fig. 8. Testing stage

Before building the bag-of-features, these testing images need to be binarized to be consistent with the image in the training model. The SIFT method is then applied to extract the hand gesture features. Because the SIFT features is invariant to scaling and rotation, the size of the testing image is not important. The next step is to employ the k-means model which to map the keypoints in the testing image into a unified dimensional histogram vector the same as the testing images. Finally, the SVM classifier model built in the training stage is employed to classify the histogram vector as one of the hand gestures trained.

3 Experimental Result

This section discusses the experimental settings and the results in details.

3.1 "Left" and "Right" Hand Gesture Recognition

There are the two experiments. The first case tries to recognize people hands which are already included in the training model. And the second case tries to recognize people hands which are not included in the training model. The training model is built with 100 training images (640×480) for the "left" and the "right" hand gestures. And the testing is performed on 100 images (640×480) to evaluate the accuracy of the multiclass SVM classifier model for each gesture.

In our experiments, we assumed that all the input images had been pre-processed with the hand gesture part of the images extracted. A low-cost Logitech QuickCam web camera provides videos captured with different resolutions such as 640×480, 320×240, and 160×120. The experiments are conducted with 200 testing images to evaluate the performance of the multiclass SVM classifier model for each gesture. In our conjecture, the bag-of-features model proposed in [1] may not perform well with hands from different people. Two experiments are built to verify our conjecture and the experiments model as show in Table 1.

Table 1. Experiment A and B (a) Training and testing both with Subject 1's hands. (b) Training with Subject 1hands and testing with Subject 2's hands

Experiment A		
Training model	Testing	Accuracy of recognition
Subject 1's hands	Subject 1's hands	95%
Experiment B		
Training model	Testing	Accuracy of recognition
Subject 1's hands	Subject 2's hands	76%

The experiment A employs Subject 1's hands to build the training model and to test with Subject 1's hands too. The experiment result achieves high classification accuracy of 95% and suggests that that the bag-of-features with the SIFT method on color/grey images can perform well for hand gesture recognition. The experiment B also employs Subject 1's hands to build the training model but employs Subject 2's hands for testing. The classification accuracy decreases to 76% from 95% as shown in Tables. The recognition accuracy of other hand gestures shows similar results. Based on the above observations, we conjecture that if all the hand gesture images are taken under the same conditions, the recognition accuracy can be improved. We propose to employ the binary images with the removal redundant hand SIFT features by different people. In Fig. 9, the color hand image on the left is binarized as shown in the left. As shown in Fig. 3, the binarized hand image retains only the shape information of the gesture and removes most of the detail information such as shadows.

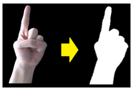

Fig. 9. Hand image after binarization

And the next experiment is to employ the SIFT method on binary images to build the training model and also in testing. We observe that by employing the SIFT method on binary images, the SIFT features extracted from different subject's hands are still similar, thus the high classification accuracy.

3.2 Five Hand Gestures Recognition

The experiments employ five hand gestures, which are C, Fist, Five, Index (point), and V (two), as shown in Fig. 10. The number of clusters to build the cluster model is an important factor that affects the classification accuracy. In this experiment, the number of clusters to build the cluster model is fixed to 800 clusters for both the proposed method and the comparing method in [1]. The experiment results are as shown in Table 2 and Table 3.

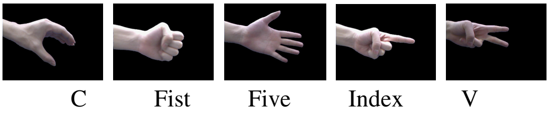

Fig. 10. Hand postures used in training images

Table 2. SIFT on color Images [1]

Gesture Name	Number of frames	Correct	Incorrect	Accuracy of recognition
C	200	200	0	100%
Fist	200	77	123	38.50%
Five	200	200	0	100%
Index	200	200	0	100%
V	200	98	102	49%
Average accuracy of recognition = 77.5%				

Table 3. Proposed SIFT on binary Images

Gesture Name	Number of frames	Correct	Incorrect	Accuracy of recognition
C	200	198	2	99%
Fist	200	196	4	98.00%
Five	200	191	9	95.50%
Index	200	193	7	96.50%
V	200	185	15	92.50%
Average accuracy of recognition = 96.3%				

In Table 2, the results show that the method proposed in [1] cannot effectively recognize two of the five gestures, namely Fist and V, but perform well to recognize the rest three gestures. Table 3, however, shows that proposed method to employ the

SIFT on binary images can recognize all five hand gesture accurately. The overall recognition accuracy is decreased when the number of gestures increases from three to five, which is as expected when employing the multi-class SVM as the classifier.

3.3 Variant Numbers of Clusters

The number of clusters to build the cluster model is an important factor that affects the qualification accuracy. The main purpose of this section is to analyze how the number of clusters affects the accuracy when using the SIFT on binary images. Table 4 shows how the number of clusters affects the classification accuracy of using SIFT on binary images. The results suggest that 1600 clusters produce the highest classification accuracy. Fig. 11 shows the comparison of the proposed method and the method in [1] with varying numbers of clusters.

Table 4. Our approach with different numbers of the clusters (codebook size)

Number of the clusters	100	200	400	800	1600	3200
Accuracy of recognition	81%	86%	92.5%	95.6%	96.3%	89.2%

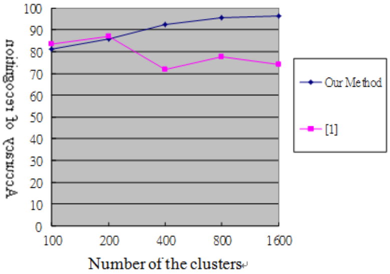

Fig. 11. Comparison with varying numbers of clusters

4 Conclusion

In this paper, we present a novel way to use SIFT on binary images for the first time in gesture recognition research. In our proposed recognition technique is inherently robust against rotations, scaling, and lighting conditions, and even with hands from different people. The proposed approach is low time-consuming approach and can provide real-time hand gesture recognition. Experiment results show that the proposed system can achieve high classification accuracy of 96.3% with hand images of different people. Three important factors affect the accuracy of the system, first is the quality of the webcam in the training and testing stages, second is the number of the training images, and third is the chosen number of clusters to build the cluster model. One of our future research directions is to apply the proposed technique for real-time sign language translation.

Acknowledgments. This work was partially supported by the National Science Council, Taiwan, under the Grants No. NSC101-2221-E-011-141, NSC100-2221-E-011-121, and NSC101-2221-E-211-011.

References

[1] Dardas, N.H., Georganas, N.D.: Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. IEEE Transaction on Instrumentation and Measurement (November 2011)

[2] Stenger, B.: Template-Based Hand Pose Recognition Using Multiple Cues. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3852, pp. 551–560. Springer, Heidelberg (2006)

[3] Tofighi, G., Monadjemi, S.A., Ghasem-Aghaee, N.: Rapid Hand Posture Recognition Using Adaptive Histogram Template of Skin and Hand Edge Contour. In: 2010 6th Iranian Machine Vision and Image Processing (MVIP) (October 2010)

[4] Kanungo, T., Mount, D.M., Netanyahu, N., Piatko, C., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Analysis and Machine Intelligence (2002)

[5] Van den Bergh, M., Van Gool, L.: Combining RGB and ToF Cameras for Real-time 3D Hand Gesture Interaction. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV) (January 2011)

[6] Xu, Z., Xiang, C., Wen-hui, W., Ji-hai, Y., Vuokko, L., Kong-qiao, W.: Hand gesture recognition and virtual game control based on 3D accelerometer and EMG sensors. In: Proceedings of the 14th International Conference on Intelligent User Interfaces (2009)

[7] Weston, J., Watkins, C.: Support vector machines for multi-class pattern recognition. In: Proceedings of European Symposium on Artificial Neural Networks, Bruges, Belgium (April 1999)

[8] Chang, C.-C., Lin, C.-J.: LIBSVM: A Library for Support Vector Machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[9] Hsu, C.-W., Lin, C.-J.: A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks (March 2002)

[10] Friedman, J.H.: Another approach to polychotomous classification. Department of Statistics and Stanford Linear Accelerator Center Stanford University (1997)

[11] Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision (2004)

[12] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (November 2004)

[13] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2006)

[14] Jiang, Y., Ngo, C., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of the ACM International Conference on Image and Video (2007)