

Analysis of the Limits of Graph-Based Object Duplicate Detection

Peter Vajda, Lutz Goldmann and Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fdrale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{peter.vajda,lutz.goldmann,touradj.ebrahimi}@epfl.ch

Abstract—Several applications require accurate and efficient object duplicate detection methods, such as automatic video and image tag propagation, video surveillance, and high level image or video search. In this paper, we explore the limits of our recently proposed graph-based object duplicate detection method. The dependency of the performance with respect to the number of training images is assessed and the optimal detection parameters are determined. Furthermore, the differences among various object classes are analyzed. In this way, this paper provides an in-depth analysis of the graph based object duplicate detection method.

Keywords-object duplicate detection, graph, SIFT

I. INTRODUCTION

A. Motivation

Image and video retrieval is an important task in computer vision. Significant efforts have been made in this area. Most existing image search and retrieval methods are based on 2D regions and features. However, these methods often fail due to changes in view points.

The general aim of object duplicate detection is to detect whether the target object is present in a set of images and to determine the locations and sizes of each occurrence. Object duplicate detection is a more specific task than object recognition since its goal is not to find all the instances of a certain object class but only a specific sample of that object class. This idea is illustrated in Figure 1 for cars and shoes.

Such an object duplicate detection approach can be useful in a number of applications. For instance, tags can be propagated to new images based on the detection of the same object in previously annotated images. Image and video search can be performed on the occurrence of a specific object, such as a suspect car in a large video surveillance database. Finally, information can be provided about an object captured by a mobile device.

B. Related work

Several research works have successfully addressed the problem of identification of specific regions in an image or video database. However, 3D object detection has not received the same interest. Therefore, in this paper, we make a step toward 3D object detection, while keeping the efficiency of 2D.

In most prior work for object duplicate detection, two specific problems can be identified. The first aims at *defining a similarity measure between image regions*. The second problem consists in *locating the position of the target object*, based on the previously defined measure.

Related to the first problem, two state-of-the-art techniques should be mentioned. The first is the "Bag of Words Model", which is based on the histogram of local features. Zhang et al. in [1] presented a comparative study on different local features on texture and object recognition tasks based on this technique. The "Bag of Words Model" does not include spatial information from the objects. However, it gives a robust, simple, and efficient solution for recognition. Conversely, the "Part based Model" considers spatial information of the local features as well. A promising method in [2] shows that the "Part based Model" performs well. More precisely, Star Model is used to represent the objects based on Histogram of Oriented Gradient (HOG) features. H. Bunke does an exhausting research on structural pattern recognition [3]. Handwriting, letter and contour line recognition is typical tasks for his works on graph based matching algorithms. In [4] the same task is performed using a 3D model of the object, where affine covariant regions are used for object modeling from video sequences.

The problem of multi-view object detection is still largely

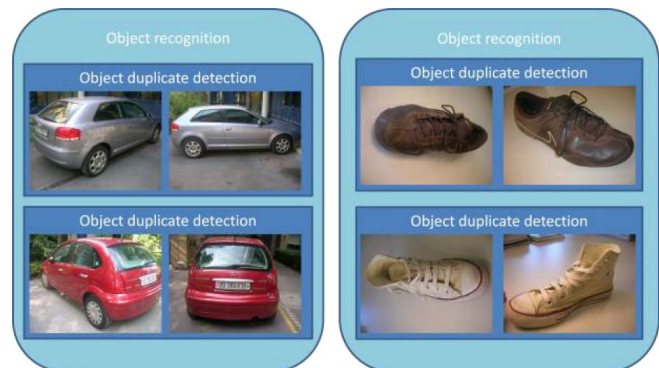


Figure 1. Illustration of the difference between object recognition and object duplicate detection. While the former groups objects into different glasses such as cars and shoes, the latter distinguishes between different shoe or car instances.

unresolved. However, some interesting solutions have been proposed for retrieving different visual views from the set of images or video. An approach described in [5] uses tracking to retrieve several different views of a same object in order to generate its representative model. The model is then used to recognize objects in a simple and accurate manner. In [5] the same task is performed using a 3D model of the object, where affine covariant regions are used for object modeling from video sequences.

For the second problem, namely, the localization of the position of the target object, affine covariant regions provide a set of points invariant to scale, rotation and translation, as well as robust to illumination changes, and changes of viewpoints [6]. On these regions, local descriptors, such as Scale Invariant Feature Transform (SIFT) [7] are extracted. The Generalized Hough Transform or a probabilistic model [8] can then be applied in order to localize the position of the object in the query image.

C. Objective

Our approach combines the advantages of being as efficient as in "Bag of Words Model", and as accurate as in "Part Based Model" due to consideration of spatial information. Another advantage of the proposed approach is that it requires a small number of training images in order to build a good model for the target object. A training phase aims at constructing the spatial relations between features in the target object, which is then represented by a graph. In other words, an attempt toward 3D modeling is made, while keeping the efficiency of 2D processing.

This work is an extension of our previous work on object duplicate detection [9]. An approach for graph-based object duplicate detection was developed and an attempt toward 3D modeling is made, while keeping the efficiency of 2D processing considering spatial information. The novelty of this paper is an in-depth analysis of the limits of this graph-based approach. 850 images were collected to evaluate our algorithm on a more comprehensive database. Optimal parameters are derived from the analysis and comparisons between various object classes are provided.

In this paper we evaluate and analyze our recently proposed object duplicate detection by answering the following questions:

- What is the difference between accuracy with different number of training images?
- What are the optimal parameter settings for certain application?
- How does the detection performance depend on the object class?

The paper is organized as follows: The method is presented in Section 2. Experiments and results are discussed in Section 3. Finally, Section 4 concludes with a summary and some perspectives for future study.

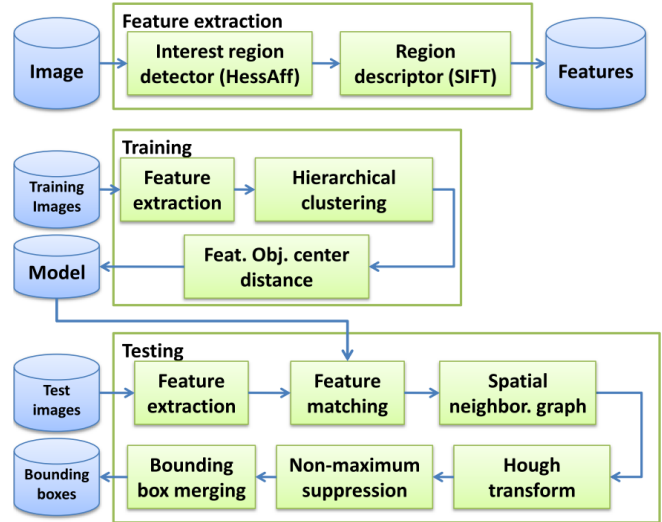


Figure 2. Overview of the object duplicate approach with the individual training and testing stages and the commonly used feature extraction.

II. OBJECT DUPLICATE DETECTION ALGORITHM

In this section, we present an efficient solution for 3D object duplicate detection in static images [9]. The goal is to detect the presence of a target object and to predict its bounding box, based on a set of images containing that object. A small number of training images, typically one to four, containing different views of the target object, are sufficient enough to achieve good performance [9].

The system architecture is illustrated in Figure 2. In the following the feature extraction, training and testing phase are explained.

A. Feature extraction

To resolve the localization problem efficiently, we use sparse features which are extracted from the training and the testing images:

- Interest regions are detected making use of a Hessian affine detector [6].
- Region descriptor extracted for each interest region. Position, scale and orientation are computed. Scale invariant image descriptors (SIFT) are then extracted from interest regions [7], as they remain robust to arbitrary changes in viewpoints.

B. Training

During the training phase, a set of images of the target object (from different views, and with eventual deformations) is processed. The training images correspond to a single object filling up the whole field of view.

- Features are extracted from the training images as described above.
- Hierarchical clustering is applied in the SIFT feature space, for improve the efficiency [10].

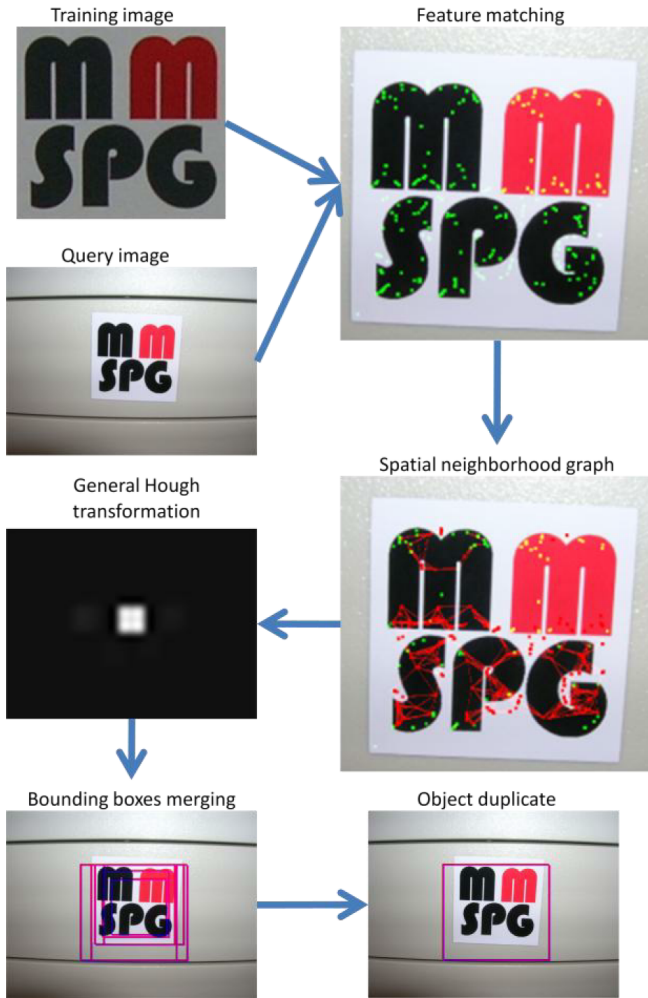


Figure 3. Illustration of the different steps of the testing stage with individual feature matching, spatial graph matching and bounding box estimation.

- Construct spatial graph model which considers the scale, orientation, position and neighborhood of the region of interest.

C. Testing

In the testing phase, we retrieve images according to a query using the following steps which are illustrated in Figure 3:

- Features are extracted from the test image as described above.
- Match features individually between the training and the testing image based on SIFT features.
- Spatial graph matching between the model from the training image and the model extracted from the test image. This considers the spatial distance of the features, hence making our algorithm more robust.
- General Hough transform to determine the bounding box of the object based on matched features.

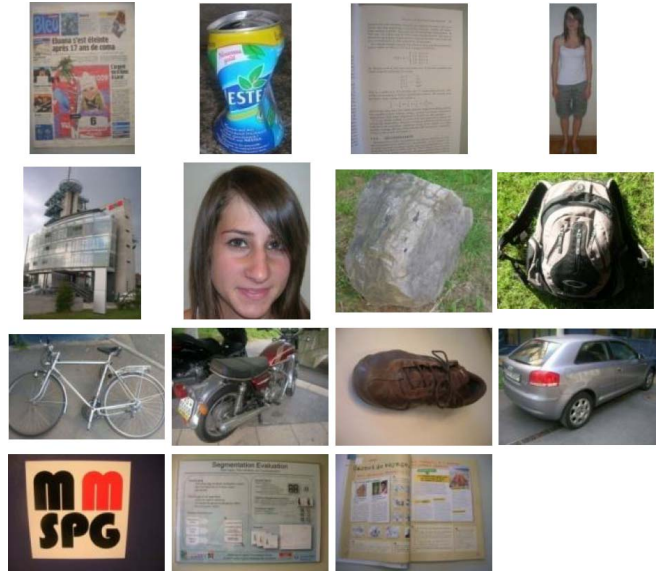


Figure 4. Samples of the different 2D and 3D object classes within the database.

- Non-maximum suppression to reject overlapping bounding boxes due to ambiguity of object location [6].
- Bounding box merging based on graph intersection to combine overlapping bounding boxes due to different views.

III. EXPERIMENTS

In this section, we assess the performance of the developed object duplicate detection method [9] on a larger dataset and explore the influence of the number of training images and the detection parameters on the performance.

A. Database

The experiments are based on a novel dataset that contains 850 pictures from 10 3D and 5 2D object classes as shown in Figure 4: *bag, bicycle, body, face, shoes, stone, can, car, building, motor, poster, logo, newspaper, book and workbook*.

Each class contains at least three samples. From each sample several images were taken from different points of view and varying distances as it is shown in Figure 5 for *building and shoes*.

B. Evaluation

Detection tasks can be evaluated based on a set of ground truth objects and a set of predicted objects which are represented by their bounding boxes. Through a pair-wise comparison of GT and PR objects a match matrix is derived that describes the correspondences between them. Based on that, a confusion matrix was created, which contains the number of true positives (TP), false positives (FP) and false negatives (FN).

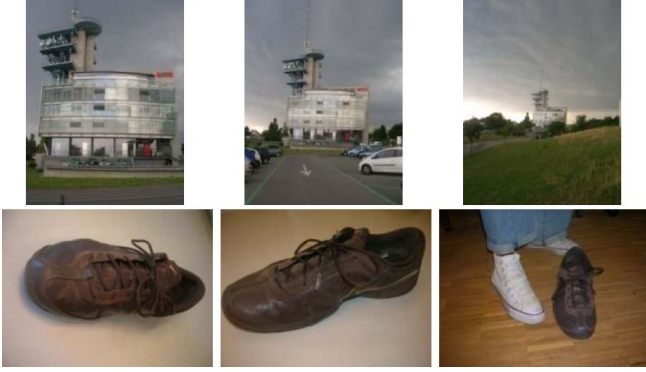


Figure 5. Samples for two objects under diverse viewing conditions within the database.

For the evaluation of detection problems two curves can be derived from this confusion matrix. The receiver operating characteristic (ROC) curve plots the true positive rate (TPR) versus the false positive per number of images (FPR), with

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN}$$

Second, precision recall (PR) curves plot the recall (R) versus the precision (P) with

$$R = \frac{TP}{TP + FN} \quad P = \frac{FP}{TP + FP}$$

not considering the TN which is not uniquely defined for a detection problems.

F-measure is calculated to determine the optimum thresholds for the object detection. It can be computed as the harmonic mean of the precision and recall values:

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad F_{\beta} = \frac{(1 + \beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

Thus it considers precision and recall equally weighted. The general formula considers the weighted values of precision and recall.

C. Results

In order to analyze the performance of the object duplicate detection depending on the number of training images, the method was trained with 1, 2 or 3 images and tested on all remaining images. Negative images (images which do not contain the ground truth object) were the different images from same class of object and several not related images. The corresponding ROC curves are shown in Figure 6. It shows a significant difference between one or more training image. The reason for that is that one training image is not enough to describe and detect an object from very different points of view. However, this result shows, more than one image can significantly improve our algorithm. In this figure we can determine, if we allow 10 FP images from 100

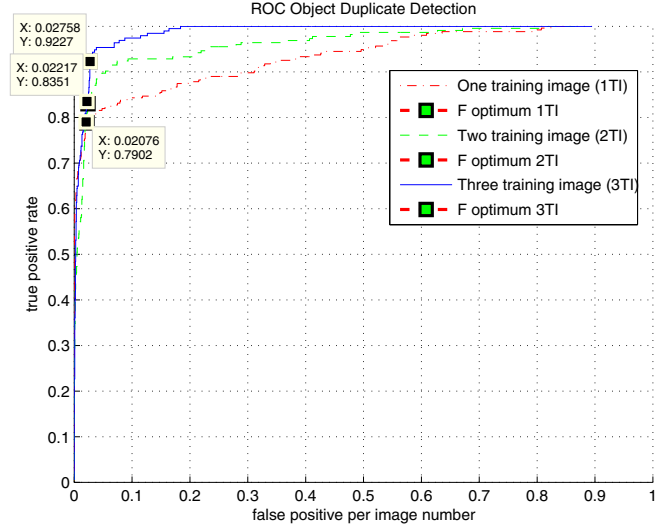


Figure 6. Receiver operating characteristic (ROC) curves for different number of training images (1,2,3) per object. A larger number of training images leads to an increased TPR for similar FP.

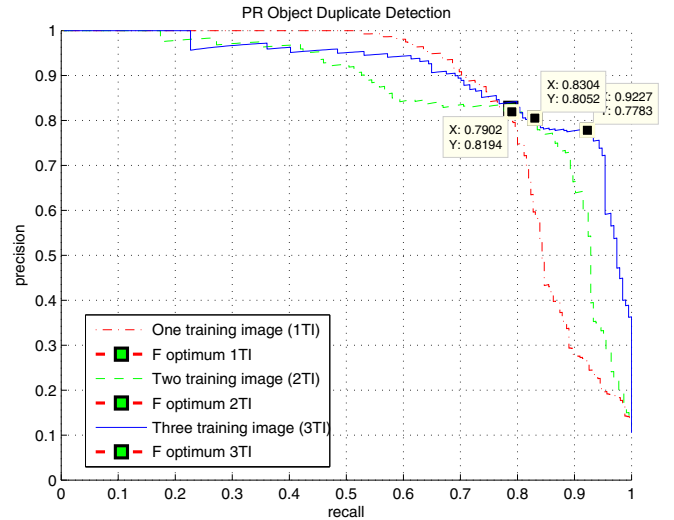


Figure 7. Recall versus precision curves for different number of training images (1,2,3) per object. Both precision and recall are increasing for more training images.

negative (which signed wrongly as contains the object), 85, 92 and 97 objects are detected, out of 100 with one, two and three training images.

The results are complemented by the PR curves shown in Figure 7, that provide a better visualization of the opposing effects (high precision vs. high recall) which are inherent to any detection task. In this figure we can determine, if at least 90 objects from 100 detected, then we detect 30, 70 and 80 right objects, out of 100 positive detected objects with one, two and three training images. If we allow the recall to be low then the precision can be lower in case of using more than one training images.

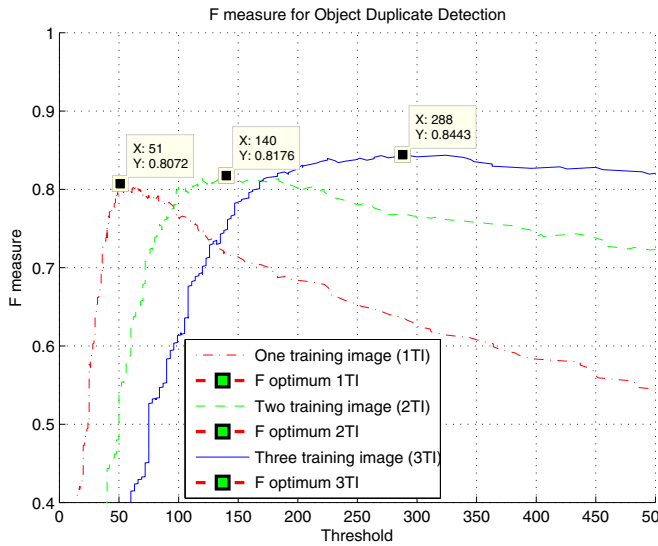


Figure 8. F-measure versus detection threshold for different number of training images (1,2,3). Both the f-measure and the detection thresholds increase with more number of training images.

In order to determine the optimal threshold for general applications, the f-measure was calculated. Two plots were shown. First, the threshold versus f-measure curve is analyzed in Figure 8. The chosen points show the maximum performance of each case. If one, two and three training images are used then an F-score of 0.80, 0.82 and 0.84 can be reached with threshold of 51, 140 and 288, respectively. Considering this curve, the optimal threshold is largely depending on the number of training images. If more training images are used the optimal threshold is increased significantly.

Second, the parameter β of the general F-measure versus threshold value is shown in Figure 9. In this plot, the importance of precision and recall can be balanced according to the requirements. If the β parameter is equal to one then precision and recall are equally important. Two other commonly used f-measures are the F_2 measure, which weights recall twice as much as precision, and the $F_{0.5}$ measure, which weights precision twice as much as recall. The optimal F-measure parameter settings are also shown in all the other figures.

In order to compare the different classes with each other, the f-measure is computed for each of the 2D and 3D object classes as shown in Figure 10. Shiny objects, such as car and motor bicycle, are hard to detect due to the changing reflections depending on the light. Books are also among the worst classes, due to large illumination variations during the image acquisition which was not the case for newspapers. Therefore, the performance of books under less diverse conditions is expected to be higher. Body duplicate detection considers just the texture of the cloths and not the shape which gives surprisingly good result. Face identification is

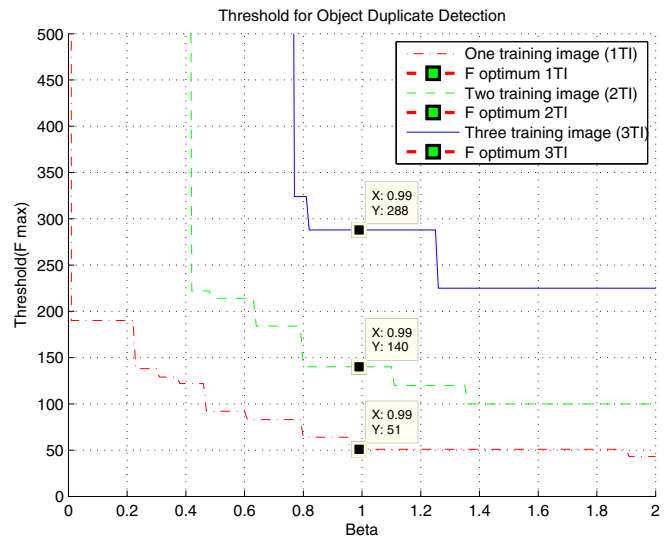


Figure 9. Detection threshold versus beta parameter of the general F-measure for different number of training images.

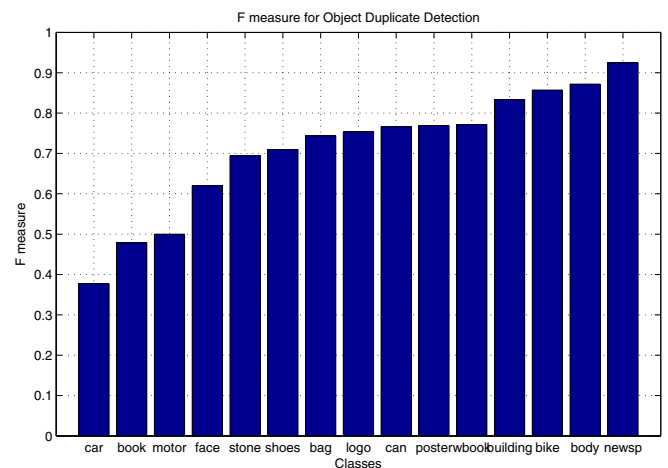


Figure 10. Performance of the object duplicate detection as f-measure across the different classes. The difference between classes is caused by different factors such as reflection properties, amount and presence of textures, amount of salient features.

also possible application; however this experiment does not consider different illuminations. Newspapers performed best, thanks to the large number of pictures and textures on the pages.

An interesting sample is discussed now in more detail. Surprisingly, even the opposite side of a car is recognized if just one training image is used. The reason is that the license plate of the car is the most salient region on both the front and the back of the car (Figure 11). Nevertheless, the location of the car is shifted upwards due to the different position of license plate with respect to the overall object.



Figure 11. Object duplicate detection algorithm detected the back side of a car, thanks to the license plate. Training image is on the bottom left corner.

IV. CONCLUSION

The recently proposed graph-based object duplicate detection algorithm was analyzed in this paper. This approach was shown to be robust when using only one or few images for training. Moreover, it was successfully tested for object duplicate detection, even when the object is captured from different views. The results show significant improvement on the accuracy if more than one image is trained from each object. Considering the F-measure and F_{β} -measure, the optimal threshold is determined. The comparison between various classes of objects shows, that shiny objects are the most difficult to detect while textured are found more easily.

As future work, we will explore the extension of this method to deal with duplicate object detection in video.

ACKNOWLEDGMENT

This work was supported by the Swiss National Science Foundation Grant "Multimedia Security", number 200020-113709, partially supported by the European Network of Excellence VISNET II (IST Contract 1-038398), and the PetaMedia (FP7/2007-2011).

REFERENCES

- [1] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, p. 213238, 2007.
- [2] P. Felzenszwalb, D. Mcallester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.
- [3] M. Neuhaus and H. Bunke, "Edit distance based kernel functions for structural pattern classification," *Pattern Recognition*, vol. 39, pp. 1852–1863, 2006.
- [4] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "Segmenting, modeling, and matching video clips containing multiple moving objects," in *Conference on Computer Vision and Pattern Recognition*, 2004, p. 914921.
- [5] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *International Journal of Computer Vision*, vol. 67, no. 2, p. 189210, 2006.
- [6] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 12, p. 4372, 2005.
- [7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, p. 91110, 2004.
- [8] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1, p. 259289, 2008.
- [9] P. Vajda, F. Dufaux, T. H. Minh, and T. Ebrahimi, "Graph-based approach for 3d object duplicate detection," in *International Workshop on Image Analysis for Multimedia Interactive Services*, 2009.
- [10] D. Nister and H. Stewenius, "Robust scalable recognition with a vocabulary tree," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, p. 21612168.