Retrieval from Image Datasets with Repetitive Structures

Praveen Dasigi and C.V. Jawahar
Center for Visual Information Technology, IIIT Hyderabad
Email:{praveend@research., jawahar@} iiit.net

***Abstract*—This work aims to enhance the matching and retrieval performance over image datasets which have similar spatial structures that occur very frequently. Instead of treating images as bags of features, we try to encode the spatial relationships in the representation. This process would help to resolve the ambiguity when two classes of images have similar sets of features although in different spatial arrangements. To demonstrate the fact a sizeable dataset of license plate images is used. We have proposed a method to use graphs to encode the spatial relationships among features. The problem of image matching thus turns to finding the maximum similarity between labelled graphs. It is shown that the precision of the retrieved results increases with this matching scheme since most of the false matches are eliminated.**

I. Introduction

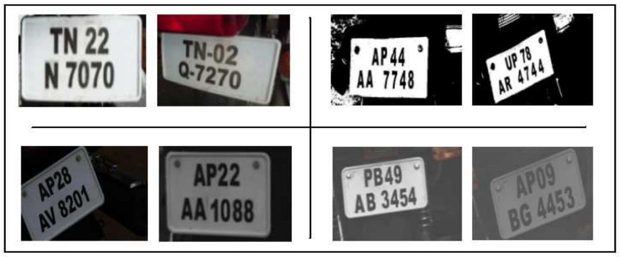
Given two images I_1 and I_2, image matching is a process that computes a score M that states how "similar" the images are. This matching score can reflect anything ranging from high level semantic similarity to low-level image similarity. Usually, a content based image retrieval (CBIR) system operates by retrieving a set of images $r_1, \ldots, r_n$ that are semantically similar to the query image Q that the user provides, based on the information need. To do so the query image has to be efficiently matched to a dataset of images $I_1, \ldots, I_N$. Any image matching process is highly dependent on the representation scheme used to encode the image structure. With the appropriate representation scheme, matching process can be improved in terms of efficiency and accuracy.

The challenge in choosing the best representation scheme for retrieval is to find the best feature sets that show invariance to various distortions in the images that occur due to different capture conditions. The feature set should provide a representation invariant to these distortions. One such popular feature set can be obtained by detecting affine invariant interest regions on the image [1]. These are regions which are supposed to be repeatable, even after significant image transformations such as projective distortions and illumination change. The feature set can be obtained by describing the detected interest regions with a SIFT descriptor [2]. The SIFT descriptor provides a 128-dimensional vector that describes the gradient orientations in the detected region by providing an orientation histogram

For retrieval, images are indexed with the feature vectors. Operating an index is faster than the process of matching the query image to every image in the dataset. Constructing an index will also allow each image to be represented with respect to the properties of the dataset. Bag of visual words [3] is a representation scheme where each feature in the dataset will be represented with a group ID or a visual codeword. The codeword is a unified representation for a group of interest regions that have similar descriptors. This is achieved by clustering the features over the whole dataset with a k-means clustering and the cluster centroids are taken as codewords. This approach is very similar to the classical vector space models for text retrieval. In the case of text, this approach considers the document as a bag of words. Though relative order may be captured, there is no point in encoding the whole geometry of document as this holds no meaning. However this aspect can be exploited for images, since, in an image the spatial arrangement of features will make clear cut sense.

This paper concentrates on a specific case of image datasets with repetitive structures. The datasets in question comprise of a very small number of unique image structures that repeat in different arrangements to form any image in the dataset. The best examples are license plate datasets or signboards where the interest regions extracted from any candidate image will most certainly be a subset of a fixed set of interest regions. If each image is considered just as a bag of visual words the relative geometry among the 'words' is neglected. For instance, in a dataset of license plates there will be many variants where the arrangement of the same set of numbers differs, which changes the semantic meaning (see Figure1). When the user queries with a licenseplate image he will be looking for "the licenseplate", rather than "a licenseplate". In such cases if retrieval happens by matching sets of words, the ambiguities due to different arrangements of same features will not be resolved. Thus, properly encoding the arrangement will enhance the precision of the retrieved set. This is achieved by using graphs to represent the image structures such that the region adjacency information is used to refine and make the matching score even meaningful.

Fig. 1. License Plates images containing repetitive structures



II. Related Work

The bag of words approach used for text retrieval has been efficiently applied to image retrieval for large object image databases and videos in particular, by Sivic et al., [4], [3]. One of the major working systems that capitalizes on the bag of visual codewords approach is VideoGoogle [3], [4]. The main proposal of this model is that image retrieval could be treated as text retrieval and similar to textual words, visual codewords are indexed. The robustness of visual codewords approach is a direct manifestation of the interest region detectors such as Harris affine detector [1] and MSER detector [5] along with the SIFT feature descriptor scheme [2]. Other descriptors that are in use to describe regions of interest are texture [6] and shape descriptors. The bag of words based retrieval system uses the concept of a visual codeword. Using visual codewords will solve this problem of low-repeatability of detected regions and improve robustness, by clustering the description vectors over the whole dataset. For the correct number of bins all the variants of interest regions that are supposed to be nearest neighbors will fall into the same bin. The bin centroid will be the visual word. This approach has been applied to the search of objects in key frames of videos, and towards object level grouping of video shots [7].

Spatial relationships among features are taken into account in [8] by applying constraints on the matches such as compactness. Quite recently some work has also gone into the area of learning the spatial relationships in [9]. In these works, attempts have been made to account for the spatial information during classification, by employing an incremental joint learning procedure to learn the shared information between regions. Graphs have been employed to pattern recognition and particularly towards shape representation by using shock trees in [10]. A taxonomy of graph applications to pattern recognition(PR) tasks has been surveyed in detail in [11] which provides a fairly up-to-date look into the area. There is a domain specific branch of literature which deals with the application of PR concepts to the recognition or classification of license plate images. Most works in this direction [12], [13] used character recognition approaches and have progressed in the direction improving the overall performance by improving the recognition accuracy at the character level. The approach that is used in this paper has been tested for a fairly large dataset of license plate images and is designed to work similarly for a class of such datasets.

III. Retrieval by Graph Matching

From the formulation in Section 3 each image is represented by a set of nodes N_i. When datasets with highly repetitive structures are used, the arrangement between nodes, if encoded will improve retrieval accuracy. **Figure 2** shows a pair of images with the same structures (characters) in different arrangements, the Bag of Words returned a similarity score of 0.79 thus making it possible to confuse them to be of same class. However using graphs, only a small part of the node edge configuration is matched and thus resulting in a more meaningful score of 0.34. However if the images belong to the same class, the matching scores from bag of words and graph matching will be almost equal.

A. Graph building and matching

Suppose a license-plate image I_i contains a set of nodes $N^i : \{v_1, \dots v_n\}$ each of them are interest regions that are extracted by a feature detector. For this we have used the MSER(Maximally stable extremal regions) [5] interest point detector. This finds a set of extremal regions that are closed under projective transformations. Following the visual-codewords approach each region will be labelled by a codeword. Thus each node v will have a label associated to it $l(v)$. Two nodes are said to be adjacent in the graph if the distance between them on the image is less than a predefined threshold. Thus an edge e_{pq} between two nodes v_p and v_q is placed if the distance criterion is satisfied. Thus for each node all the nodes within a certain radius will be its neighbors. Thus the graph G_i for an image I_i is stored as an adjacency matrix along with the visual codewords from the image. In the building phase the variable 'kd' which specifies the adjacency threshold for placing the edges. For smaller values of kd only small local regions are identified and the graph is very sparse. The most optimal value of kd that models the necessary set of relationships has to be experimentally found out.

In the general form of a license-plate retrieval system, license plate images are captured from varying views and illumination conditions and thus the resulting images are reasonably distorted from one another. The job of the retrieval system is to retrieve similar images given a query image. In a parts based setup as being described above. The MSER detector which detects the parts will be able to output regions which are invariant to the distortions to a certain extent. However the regions are not 100% repeatable. Hence there is no guarantee that the graphs formed from two images of the same license plate will be exactly similar. Hence the matching scheme should be able to take care of this aspect.

Given two graphs G_i and G_j, the first possible method to match them is to find if both the graphs are isomorphic and output a $0/1$ similarity score. This would be exact graph matching when G_i is a exactly similar to G_j or is exactly a part of G_j. In complexity literature this is an NP complete problem and some approximate solutions have been found [11] to break the complexity. Due to low repeatability of detectors when there is discrepancy in even one of the nodes, an exact graph matching algorithm will output a 0. Our need is to find the amount of similarity between the graphs. The same problem applies using graph-subgraph isomorphism as well. The next possible approach is to find the maximum common subgraph(MCS). Given two graphs G_1 and G_2, the MCS is the graph g that is common to both graphs in terms of node and edge configurations. To find MCS all possible permutations are to be evaluated . Effective steps have been proposed to break the complexity of this problem in [14]. Using MCS will solve the problem to a certain extent as there is more flexibility to the size of the subgraph and the graphs could be of any sizes. However the problem of low repeatability still haunts, as even the parts of the graphs which are almost similar will be left out. Judging from these aspects it is imperative that the exactness constraint has to be relaxed. Thus the requirement is that the algorithm has to find the best subgraphs in G_1 and G_2 that are as similar as possible and a score has to be given that depicts the similarity of these subgraphs.

Thus, given two graphs with node and edge configurations $\{G_1 : (N_1, E_1)\}$ and $\{G_2 : (N_2, E_2)\}$ the goal is to find the best mapping out of all mappings $m \subseteq N_1 \times N_2$ such that the similarity score $sim(G_1, G_2)$

$$sim(G_1, G_2) =$$
$$max_{m \subseteq N_1 \times N_2}\{s(\{N_1, E_1\} \sqcap_m \{N_2, E_2\} - d(reps))\}$$

Here sim is the maximum similarity between graphs m is the best mapping which maximizes the score greedily. s is the score computation function for the mapping and d is the discounting function for taking care of one to many matches in the mappings.

B. Maximum graph similarity

The problem of finding the similarity between graphs to eliminate ambiguities of image match is shown in Figure 2. To find the Efficient similarity, the algorithm builds the set of mappings between N^i and N^j. The mapping $m(v_i)$ are a set of nodes v_j in G_j that may be possibly matched to a node v_i in G_i. In this case, this mapping function $m(v_i)$ will find all the nodes with the same codeword-label. If no such labels are found the top-k nodes with least L1 distance between the SIFT descriptors are found out. After each of the mappings are found, the process is to evaluate all the combinations of mappings for the best overall similarity score such that this can be used as the maximum similarity between the two graphs. This would find the global maxima of the scores. However this process is also NP-complete. This is because, though all the combinations of all the mappings, will be less than the entire search space i.e., $2^{(|N^i| \times |N^j|)}$, since the mappings are for a reduced number of nodes, even then it will be highly expensive. Thus we adopt an approximate solution by the use of greedy evaluation of mappings. This is outlined briefly in the Algorithm **1**. This is a variant of the algorithm proposed for comparing CAD diagrams in [15].

To find the maximum similarity, each of the codeword-matches and their neighbors are evaluated. For every new match, the pair is eliminated from the search. At each level the match with best score for the node in G_i to the set of remaining nodes in G_2 is evaluated. This greedy approach may not be able to find the most optimal solution. The matched sets and the similarity score will be sub optimal. It runs in polynomial time i.e, $O((|N^i| \times |N^j|)^2)$

IV. Experiments and Results

A. Experiments

Setup: The retrieval experiments will be explained in this section. For retrieval we have built a dataset of license plates with the following properties. The dataset has 5000 images of 250 vehicles at 20 angles of capture. This license plates are entirely random and taken at different times of day at different locations, thus resulting in different illuminations and backgrounds. The retrieval experiment comprised of querying the system with a cropped version of an image that shows only a license plate and the system is expected to retrieve all the images that match this image in terms of having these components in the image. Three matching schemes are used for the retrieval process. They are described briefly here.

NDM: In this process the descriptors that are computed from the affine invariant regions are directly used for matching as in [2]. An L1 distance is computed between the 128-vectors

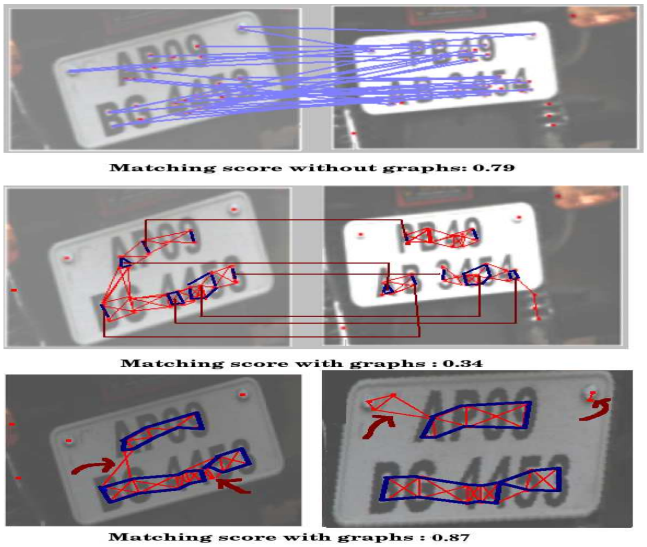

Matching score without graphs: 0.79
Matching score with graphs : 0.34
Matching score with graphs : 0.87

Fig. 2. Similarity scores with and without graphs, first figure shows bag of words match on semantically different number plates the matching score is higher even though the plates are different, second figure shows the match is refined with graphs since arrangement is taken into account, in the third figure most of the graph is matched since the images are only view distorted variants of the same plate. Figure serves as proof of concept that using graphs will improve discriminability

Algorithm 1 Calculate MaxSim from all mappings

```
$simmat \Leftarrow$ similarity matrix between all nodes $v_1 \times v_2$
$bestpairs \Leftarrow 0$
for $i$ in all matches $m$ do
  for $j_a, j_b$ in all pairs of match $i$ do
    $nb_1, nb_2 \Leftarrow$ immediate neighbors of $j_a, j_b$
    Find best match $bm_1$, $nb_1(k)$ in $nb_2$; retain only global best
    Add $\{j, bm_1\}$ to bestmatches,Eliminate matched pairs
    Find best match $bm_2$ for $nb_1(k)$ in $nb_2$
    Add $\{j, bm_2\}$ to bestmatches,Eliminate matched pairs
    Retain best matched pair out of all repetitions
  end for
  $scr(i) \Leftarrow$ avg similarities of pairs in $bestmatches$
end for
$MaxSim \Leftarrow mean(scr)$
```

and a similarity measure s computed from the distance. This is the most standard comparison scheme. The performance of this method is affected by the repeatability of the descriptors, the overall retrieval performance is demonstrated in Figure 3(a).

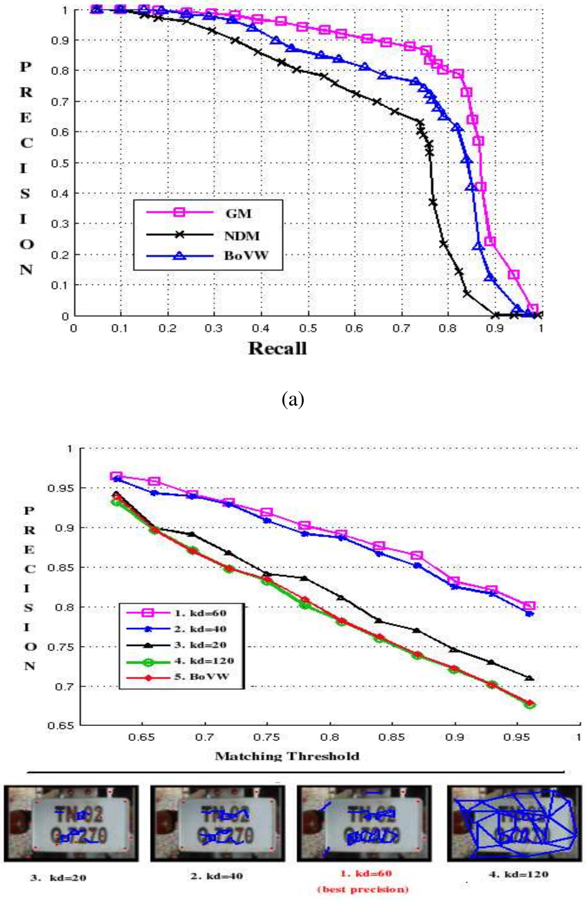
BOVW: In the bag of visual words (BoVW) [3] approach the descriptor vectors that are computed for the all the regions of every image in the dataset are clustered and the bin centroids are taken as visual words. Matching between two images proceeds by computing the amount of similarity between the two sets. Since the problem with the low-repeatability of the descriptors is taken care by the codeword representation the retrieval performance is much better than the NDM scheme. This fact is demonstrated in the precision recall curves in Figure 3(a).

GM: In the graph matching based approach, graphs are built on each of the images in the dataset and the matching score is computed as described in the Algorithm 1. Using this scheme has increased the retrieval precision in cases where the dataset contains images that are ambiguous. These ambiguities are basically images that have the same characters on the license plate but are shuffled in some order, obviously differing by semantic meaning. Since this approach also takes care of the spatial relationships, between words, the top retrieved images will be almost error-free which can be seen in detail in Figure 4.

B. Results

The precision recall curves shown in Figure 3(a) illustrate the facts. The curves are plotted as follows. Retrieval is performed on the dataset using the three approaches as detailed above. For each of the methods used, the average precision values are plotted against corresponding recall values varied by the similarity threshold from 0.0 to 1.0 . It can be seen that the NDM approach performs quite poorly when compared to the other two. The precision is fairly low which means that for a distance threshold many false-matches images are retrieved. This is expected due to the low-repeatability of descriptors and sensitivity to descriptor matching score.These problems are alleviated in BoVW scheme since the codewords account for all the variations of the regions clustered into the corresponding bin. Thus there are no false matches due to the ambiguity of the descriptor matching score. The precision is thus improved over the NDM scheme. However since no spatial relationships are encoded, some false matches owing to the ambiguity in the order of detected regions occur. In the GM scheme, graphs are used to encode spatial arrangement. Thus false matches are eliminated as the shuffle between detected regions is discounted in the matching score between two graphs as the corresponding edges do not match. Thus GM is mainly a refinement over BoVW.

In Figure 3(b), the precision values are plotted for similarity score thresholds ranging from 0.65 to 0.95. Each graph shows the precision values for different nearest neighbor distance ‘kd’ used to place edges along with the precision values for BoVW based match. For a smaller kd=20, the precision values are quite near to the BoVW scheme as edges are placed between very close regions only and there is no overall improvement in resolving the ambiguity.



(a)

(b)

Fig. 3. (a): PR curves with NDM, BOVW and GM retrieval schemes. (b): Precision values at various matching thresholds, for graph NN-distances kd=20,40,60 and 120, and for BoVW scheme. Notice that for kd=120 its almost a complete graph and the performance is similar to BoVW since the graph doesn’t resolve any ambiguity. kd=60 results in best precision and kd=20 lies midway as edges represent only a small amount of relationships

The same happens for kd=120 as edges are placed between almost every pair of regions and the graph is almost a complete graph. In this setup the precision values are almost equal to BoVW as every region will be adjacent to every other and no new information is added. For kd=60 the best precision occurs which means that the arrangement is both necessary and sufficient. Thus the graphs are best built with NN-distance threshold to be 60. At kd=40 the precision is slightly less though better than others.

The retrieval results are shown for a sample query of a licenseplate image in Figure 4. As the marked regions of the image indicate, graph based retrieval is able to retrieve the best ten matches for the query images. The other two approaches i.e., NDM and BoVW have retrieved 4 and 2 false matches due to the problems in repeatability and sensitivity in the first

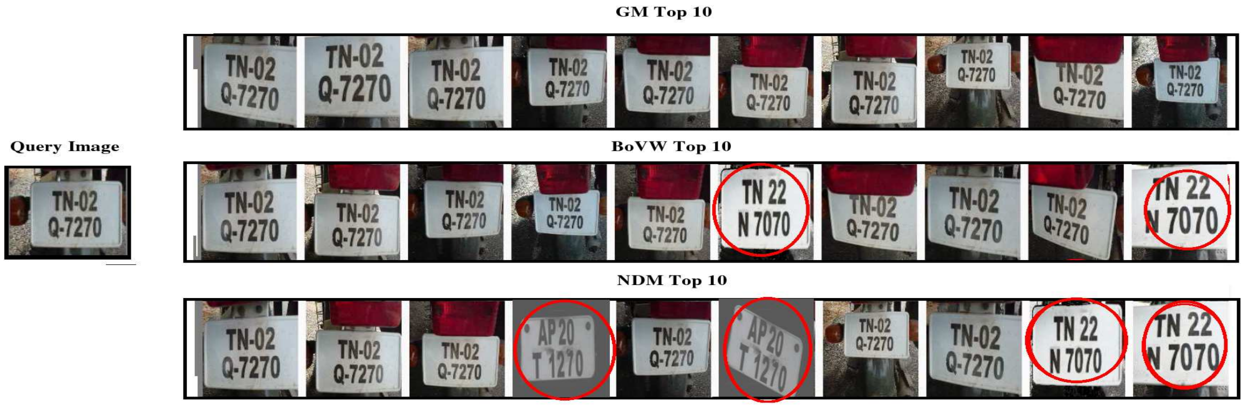


Fig. 4. Query and retrieval results with GM, BOVW and NDW for a sample query. The false results are marked in red-circles.

case and lack of spatial information in the second case. The results marked in red show the images which are false matches that returned similarity scores within the threshold

V. Conclusions and Future Work

The most logical extension for this work is to increase performance in both the computational and the scalability levels. Right now, scaling the dataset to reasonably large size is an issue because of the computational requirements of graph matching. Thus it would be a good result if the graphs are represented in a lower dimension viable for both indexing and retrieval complexity. Since the focus is on datasets with repetitive structures, there is scope for learning the structures that contribute to matching and eliminating the non-contributing regions. This would improve the precision as well as the computational time. One other direction is to employ a graph indexing scheme that indexes the most representative graph substructure such that retrieval will become much faster.

References

[1] K. Mikolajczyk and C. Schmid, “Scale and Affine Invariant Interest Point Detectors,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[2] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. of the International Conference on Computer Vision ICCV, Corfu*, 1999, pp. 1150–1157.

[3] J. Sivic and A. Zisserman., “Video google: A text retrieval approach to object matching in videos,” in *Proc. of the International Conference on Computer Vision*, October 2003, pp. II:1470–1477.

[4] J. Sivic, F. Schaffalitzky, and A. Zisserman., “Efficient object retrieval from videos,” in *Proc. of the 12th European Signal Processing Conference EUSIPCO 04, Vienna, Austria,*, September 2004.

[5] J. Matas, O. Chum, U. Martin, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” in *Proceedings of the British Machine Vision Conference*, vol. 1, London, 2002, pp. 384–393.

[6] M. Varma and A. Zisserman, “Statistical approaches to material classification,” in *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec. 2002, pp. 167–172. [Online]. Available: http://www.robots.ox.ac.uk/ vgg

[7] J. Sivic, F. Schaffalitzky, and A. Zisserman., “Object level grouping for video shots,” in *Proc. of the 8th European Conference on Computer Vision, ECCV Prague, Czech Republic Springer Verlag*, May 2004, pp. II:85–98.

[8] F. Schaffalitzky and A. Zisserman, “Automated location matching in movies,” *Computer Vision and Image Understanding*, vol. 92, pp. 236–264, 2003. [Online]. Available: http://www.robots.ox.ac.uk/ vgg

[9] A. Opelt, A. Pinz, and A. Zisserman, “Incremental learning of object detectors using a visual shape alphabet,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2006, pp. I:3–10.

[10] A. Shokoufandeh, L. Bretzner, D. Macrini, M. F. Demirci, C. Jönsson, and S. Dickinson, “The representation and matching of categorical shape,” *Comput. Vis. Image Underst.*, vol. 103, no. 2, pp. 139–154, 2006.

[11] D. Conte, P. Foggia, C. Sansone, and M. Vento, “Thirty years of graph matching in pattern recognition,” in *Proc. of the Intl. Jnl. of Pattern Reconition and Artificial Intelligence IJPRAI*, 2004, pp. 265–298.

[12] A. Mecocci and C. Tommaso, “Generative models for license plate recognition by using a limited number of training samples,” in *In Proc. of International Conf. on Image Processing (ICIP)*, 2006, pp. 2769–2772.

[13] M. Donoser, C. Arth, and H. Bischof, “Detecting, tracking and recognizing license plates,” in *In Proc. of Asian Conference on Computer Vision (ACCV)*, 2007, pp. 447–456.

[14] H. Bunke, P. Foggia, C. Guidobaldi, C. Sansone, and M. Vento, “A comparison of algorithms for maximum common subgraph on randomly connected graphs,” in *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*. London, UK: Springer-Verlag, 2002, pp. 123–132.

[15] Pierre-Antoine and C. Christine, “Measuring the similarity of labeled graphs,” in *Proc. of International Conference on Case-Based Reasoning*, 2003, pp. 80–95.