Car Make and Model recognition combining global and local cues

Meena AbdelMaseeh, Islam Badreldin, Mohamed F. Abdelkader and Motaz El Saban
Microsoft Research, Advanced Tehcnology Labs in Cairo, Egypt
motazel@microsoft.com

Abstract

This paper addresses the problem of Car Make and Model recognition as an example of within-category object class recognition. In this problem, it is assumed that the general category of the object is given and the goal is to recognize the object class within the same category. As compared to general object recognition, this problem is more challenging because the variations among classes within the same category are subtle, mostly dominated by the category overall characteristics, and easily missed due to pose and illumination variations. Therefore, this specific problem may not be effectively addressed using generic object recognition approaches. In this paper, we propose a new approach to address this specific problem by combining global and local information and utilizing discriminative information labeled by a human expert. We validate our approach through experiments on recognizing the make and model of sedan cars from single view images.

1 Introduction

The problem of object recognition captures a lot of interest in the computer vision community. This term usually refers to a whole set of problems of finding an object in a scene and classifying it into one of a set of general broad categories representing the generic nature of the object such as vehicles, mugs, or humans. Hence, the main role of object recognition approaches is to build a general model for a category that captures the overall global characteristics of that category as compared to other categories. For example, a perfect recognition technique would learn how a sedan car would appear, regardless of the specific car make and model.

On the other hand, many applications require the finer level of object identity recognition, such as car make and model recognition. In such applications, the global category of the object is given, and it is required to recognize the exact object class among similar objects within the same category. We refer to this problem as within-category object class recognition. The generic object recognition approaches may appear to be applicable to within-category scenarios. However, the performance of these approaches fall down considerably in this case. These failure cases are expected, since the variations between the within-category objects are often subtle and minor as compared to variations induced by clutter and illumination changes.

In order to overcome these challenges, we propose in this paper a novel approach for within-category object class recognition. The first step of our approach is to extract visually discriminative parts from the query images corresponding to human-annotated parts in the gallery images. The correspondence is found using global shape descriptors, since the within-category classes share the same global shape and part configurations. The second step is to describe the extracted discriminative parts using local shape and appearance descriptors. Finally, the final matching score to a gallery image is found by a weighted sum of the global and local scores. We validate our approach through experiments on recognizing the make and model of sedan cars from single view images.

2 Related Work

Despite the overwhelming amount of work on global object recognition, few papers [1, 2] have investigated the more specific problem of within category object recognition. In [1], the problem is investigated in the domain of subordinate-level categorization in the avian domain. A volumetric framework is used in combination with an appearance model. Another interesting form of the problem is investigated in [2], where the authors refer to the problem as visual identification and attempt to learn discriminative appearance patches.

In the car recognition domain, few attempted the finer classification problem of identifying car make and model. However, all of these approaches have treated the problem in a manner similar to the general object

categorization. For example, [3] attempted to directly identify car make and model from projected images of a 3D CAD model. The matching is done based on features similar to SIFT features but with rotation invariance deliberately disabled. [4] also attempted to identify car make and model but from frontal images only. Different types of features are extracted from pre-defined regions and compared to the database cars. A shape based approach was presented in [5], where cars are identified from back views only by using features extracted from the car backlights and measurements from the car global shape. More recently, [6] attempted to tackle the problem of within-class car recognition by combining SURF features and bag-of-words model with structural verification techniques and validated their approach on realistic-looking toy car datasets.

3 Combining Global and local descriptors

The algorithm starts with an edge map sketch of a gallery image, a group of points uniformly spanning the edge maps are selected. For each selected edge point, a global shape descriptor is computed. Another local shape descriptor is computed for edge points belonging to manually-annotated local parts. In addition, appearance features and their descriptors are extracted from the manually-segmented regions in each gallery image. Those parts and regions, annotated by a human expert, are visually discriminative across different object classes of the same category. Finally, the appearance descriptors are indexed in a KD-tree.

In the query phase, an edge map is first computed using probabilistic boundary detector [7]. The global shape descriptors of all edge points are computed and then used to perform a global registration of the query image to each of the templates, and compute a global dissimilarity measure. We also use the registration correspondences to extract the corresponding parts to the manually-annotated discriminative parts of the gallery images. Local shape and appearance descriptors are computed for each of these parts, and matched to the corresponding local descriptors in the gallery images. The query is assigned to the class with the minimum weighted sum of the global and local dissimilarity measures.

3.1 Global shape context descriptor

Shape context (SC) [8] is an effective shape descriptor for global shape description and finding correspondence between two set of points belonging to different objects. The SC descriptor h_i at a given object point $p_i \in \{p_1...p_K\}$ describes the distribution h_i of other object points relative to it. The cost of matching p_i to $q_j \in \{q_1...q_K\}$ belonging to another object is given by: $c(p_i, q_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k)-h_j(k)]^2}{h_i(k)+h_j(k)}$, while the correspondence between the two sets of points can be formulated as an optimal assignment problem $q_j = \pi(p_i)$. This assignment attempts to minimize the global dissimilarity measure defined by:

$$C_g = \sum_{i=1}^{K} c(p_i, q_{\pi(i)}). \quad (1)$$

Due to the global shape similarity between all the within-category objects, equivalent points of objects of the same category tend to have close SC descriptors leading to robust registration even in the presence of outliers. However, the dissimilarity measure C_g in most cases cannot discriminate between similar looking inter-class objects, mainly due to the coarse quantization and the log scaling of h.

3.2 Local shape descriptor

Consider an arbitrary sequence of points $p^* = \{p_1^*...p_S^*\}$ that belongs to a contour labelled by a human annotator to be visually discriminative across different instances of the same class. Using SC matching, a corresponding sequence $\hat{q}^* = \{\hat{q}_1^*...\hat{q}_S^*\}$ is estimated. We chose not to use histograms to describe the shape of local contours to achieve more detailed and uniform description. Instead, two matrices similar to those described in [9] are computed for each of the two sequences. The first matrix D represents all pairwise distances, while the second Θ represents all pairwise orientations among different points of the same sequence. Pairwise distances and orientations are defined as $D^{p^*}(i,j) = \frac{S^2\|p_i^*-p_j^*\|_2}{\sum_{i=1}^{S}\sum_{j=1}^{S}(D^{p^*}(i,j))}$ and $\Theta^{p^*}(i,j) = \frac{2\angle(p_i^*-p_j^*)}{\pi} \in [-\pi, \pi]$ respectively. Two local dissimilarity measures can be then estimated using the sum over pairwise distances.

3.3 Local appearance matching

While local shape matching can capture some distinctive local features such as the shape of the contour of the rear window, it fails to capture the distinctive appearance features of some car parts such as the backlights or the make logo. Hence, we combine local shape matching with local appearance matching to capture most of the distinctive local parts information.

3.3.1 Parts Segmentation

Since we need to find interest points in the local parts only of the query image, such as the backlights, we first segment the relevant part out of the image. A convex polygon can be estimated to include all contour points of $\hat{q}^* = \{\hat{q}_1^* ... \hat{q}_S^*\}$, provided that $S \geq 3$ and the points are not collinear. This convex polygon is then used as a segmentation mask to black-out all the query image except for the relevant local part. This masked image is then utilized for feature extraction as described next.

3.3.2 Appearance Feature Points Detection, Description, and Matching

Given an image of a local part, we detect a set of keypoints $K = \{K_1, K_2, \ldots, K_L\}$ using the FAST detector [10]. Then, these points are described using one of the DAISY decriptors [11]. We choose to use the T2_8a_2r6s_PCA32 descriptor that gives a good blend of speed, low error rate, and low dimensions. For feature matching, the state-of-the-art approach is to match the set of keypoints K to a database of features using a K-nearest neighbor (NN) approach with an efficient indexing scheme such as a KD-tree, while using the NN ratio for discarding outliers as in [12]. For this purpose, after feature extraction and description of the gallery images, all the feature descriptors are stored in one KD-tree corresponding to the local part (e.g. backlights).

The matching between a query image and the top M matches of the gallery images is then done using a voting scheme. Consider the j^{th} keypoint K_j in the query image, the nearest neighbors in the gallery keypoints macthing K_j, in terms of Euclidean distance, are identified using a *ratio test* procedure similar to [12]. For each keypoint P_i in the gallery, the Euclidean distance between the descriptor of P_i and the descriptor of K_j is calculated and then normalized by the *outlier distance*. Due to the shell property of high dimensional spaces, the non-matching points in the gallery images all tend to lie approximately the same descriptor space distance away from the query point K_j. This distance is called the outlier distance. Using the KD-tree, the outlier distance can be found as the distance to the k^{th} nearest neighbor. The acceptance of a point P_i can then be done based on a *ratio test*, where the ratio, is defined as

$$\text{Ratio}(K_j, P_i) = \frac{\|\text{Desc}(K_j) - \text{Desc}(P_j)\|}{\text{outlier distance}} \qquad (2)$$

where $\text{Desc}(X)$ is the descriptor of a feature point X. The *ratio test* then accepts all the points P_i that have a ratio, as defined in equation (2), that is higher than a certain threshold (the ratio test threshold parameter, R_0). We can then define the *strength* of a match as the reciprocal of the ratio defined in equation (2). Our voting scheme then ranks the individual gallery images by aggregating the strength scores of their individual keypoint matches. Finally, the dissimilarity measure C_A for appearance matching for a gallery image is taken as the reciprocal of the sum of the strengths of the feature match pairs belonging to that gallery image, where the match strength is the reciprocal of the ratio defined in equation (2). Hence,

$$C_A = \Big(\sum_{\substack{P_i \in P,\, K_j \in K \\ \text{Ratio}(K_j, P_i) > R_0}} \frac{1}{\text{Ratio}(K_j, P_i)} \Big)^{-1} \qquad (3)$$

where P is the set of feature points in a gallery image, K is the set of feature points in a query image, and R_0 is the ratio test threshold parameter.

4 Experimental results

We performed experiments on detecting the make and model of the 'car' category of the Savarese et al. 3D object recognition dataset [13]. We opted for this dataset because it contains 10 different car makes and models of Sedan type at different scales, poses, lighting conditions. Only the rear view was investigated in our experiments (total of ten cars each having three different images)[1], and to exclude the effect of background clutter, we manually segmented the car images in this dataset from the background. Template descriptors are built by collecting example images of the 10 different car makes and models in this dataset from the Internet. We manually cleaned up one image for each class and used it for extracting edge map sketches used for global shape descriptors. We manually annotated the left backlight of each template and used it as a discriminative part. A local descriptor for this part in each class is computed using the sketch edge points, and the part images for all the classes constituted the gallery from which appearance features were extracted and stored in the parts KD-trees.

Over all the query images, the segmentation precision was **62.53%** and the recall was **79.09%** compared to manual segmentation. This proves the success of the SC global shape descriptor in establishing correspondence between query and template discriminative parts. Correct classification rates using different dissimilarity measures are shown in Table 1. It can be seen that our proposed approach improves the correct classification rate over the cases where global or local descriptors

[1] It is worth noting that the rear view was chosen as an example view to serve as a proof for the concept of combining local and global cues.

Desc	Global SC (C_g)	Local angular (C_Θ)	Local distance (C_D)	Combining local and global (C_g, C_D, C_Θ)	Appearance (C_A)	Our approach (C_g, C_D, C_Θ, C_A)
Top-1	43.3%	33.3%	26.7%	36.7%	23.3%	**53.3%**
Top-2	53.3%	53.3%	46.7%	53.3%	40%	**60%**
Top-3	56.7%	53.3%	56.7%	76.7%	50%	**70%**

Table 1. Top matches correct classification rates for different similarity measures.

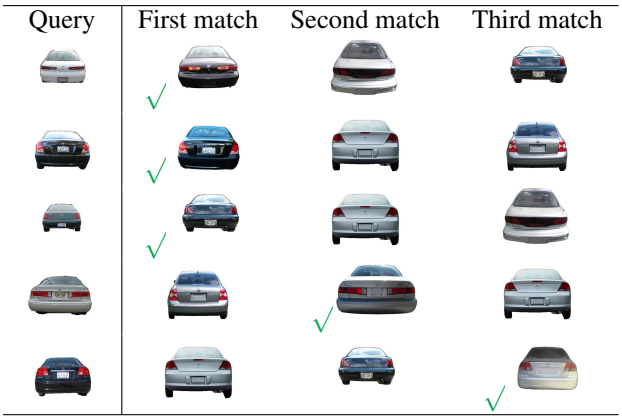

Figure 1. Examples of top-3 matches from our approach.

alone are used. Some examples of our classification results are shown in Figure 1.

5 Conclusion and Future Work

In this paper we proposed a new method for within-category fine grained object class recognition by combining both global and local object information. We investigated the effectiveness of our approach in the problem of detecting the make and model of sedan cars. Though the proposed approach achieved satisfactory performance results, it would be interesting to test the the approach with a larger dataset and to measure performance for different poses, and with other object categories.

References

[1] R. Farrell, O. Oza, N. Zhang, V. Morariu, T. Darrell, and L. Davis, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in *Computer Vision, 2011. ICCV 2011. IEEE 13th International Conference on*, Nov. 2011.

[2] A. Ferencz, E. Learned-Miller, and J. Malik, "Learning to locate informative features for visual identification," *International Journal of Computer Vision*, vol. 77, pp. 3–24, 2008.

[3] J. Prokaj and G. Medioni, "3-D model based vehicle recognition," in *Applications of Computer Vision (WACV), 2009 Workshop on*, Dec. 2009, pp. 1–7.

[4] V. Petrovic and T. Cootes, "Analysis of features for rigid structure vehicle type recognition," in *In British Machine Vision Conference*, 2004, pp. 587–596.

[5] D. Santos and P. Correia, "Car recognition based on back lights and rear view features," in *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09. 10th Workshop on*, May 2009, pp. 137–140.

[6] D. Jang and M. Turk, "Car-Rec: A real time car recognition system," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, Jan. 2011, pp. 599–605.

[7] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 5, pp. 530 –549, may 2004.

[8] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 4, pp. 509 –522, Apr. 2002.

[9] T. Ma and L. Latecki, "From partial shape matching through local deformation to robust global shape similarity for object detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, Jun. 2011, pp. 1441 –1448.

[10] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 105–119, Jan. 2010.

[11] S. Winder, G. Hua, and M. Brown, "Picking the best daisy," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, Jun. 2009, pp. 178–185.

[12] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[13] S. Savarese and F. Li, "3D generic object categorization, localization and pose estimation," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, Oct. 2007, pp. 1–8.