16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

CATEGORY-LEVEL DETECTION BASED ON OBJECT STRUCTURES

Alex Y.S. Chia[1]*, Deepu Rajan*[1]*, Maylor K.H. Leung*[1]*, and Susanto Rahardja*[2]

[1]Nanyang Technological University
Nanyang Avenue, Singapore 639798

[2]Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

We present a new class of descriptors which exhibit the ability to yield meaningful structural description of the object. These descriptors are constructed by harnessing the geometrical relationships and spatial configurations between two types of image primitives: Quadrangles and ellipses. Specifically, we extract the line segments from the line edge map of the image and exploit the spatial qualities of the line segments and the salient colors of the image to construct the quadrangles. The ellipses are extracted with a close loop system that is driven by Gestalt Psychology. Experimental results show very good performance for category-level object detection in which the objects in each category exhibit variations in form, scale and viewpoint.

1. INTRODUCTION

Detecting objects belonging to the same category in real world scenes has proven to be a difficult challenge for computer vision. Given the vastly superior performance of the human in categorizing multiple classes of objects, we turn to human psychology for inspirations. In particular, we recognize that the human can easily detect different objects belonging to the same category based on the structures/shapes of the objects. In this regard, descriptors that closely model the structures of the objects may be useful for category-level object detection.

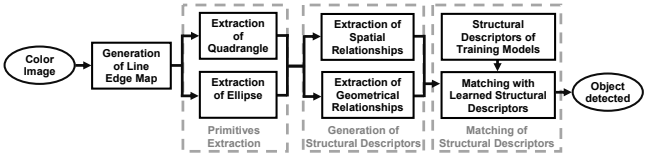


Figure 1: System Flowchart.

Our main contribution in this paper is a set of structural descriptors that exhibit the ability to yield meaningful description of the object. These descriptors are constructed from two types of image primitives: Quadrangles and ellipses. We show the flowchart to generating the descriptors in Fig. 1. We extract the image primitives from the line edge map of the image. Each pair of line segments defines a region of space that is enclosed by the pair of line segments and the end points of their overlap portions. We term these regions as the quadrangles. These quadrangles can be interpreted as the building blocks for the 2D counterparts of the generalized cylinders which have been demonstrated to yield meaningful structural description of complex 3D objects [1–3]. Given that a quadrangle is composed by a pair of *straight* line segments, it is not suited to model objects whose shapes are predominantly compose by *curved* segments. To this end, we model the curved segments with the boundary of an ellipse. We choose an ellipse since it is one of the simplest circular shape. In addition, the line drawings of many man-made objects often exhibit instances of an ellipse [4]. We exploit the geometrical and spatial relationships between the primitives to construct the structural descriptors. Finally, by matching these descriptors with those previously extracted from the training images, we detect the objects present in the image.

1.1 Related Work

Early works on using shape information for object recognition extract features that characterized the *global* shape of an object e.g. the Fourier transform [5, 6] and the medial axis transform [7]. Although global features can be extracted at low costs and are effective for certain classification tasks, these features are sensitive to occlusion and noise. As such, their success in real world scene is limited.

Recently, Belongie *et al.* [8] developed the semi-local edge features "Shape Context" which recorded the statistics of the edges into log-polar histogram bins. A similar approach was proposed by Carmichael and Hebert [9] who characterized an edge point by the distribution of the edges in an aperture surrounding the edge point. Similarly, Mikolajczyk *et al.* [10] used the positions and orientations of the edges to construct scale-invariant shape descriptors. We highlight that the edge features extracted from these methods consider each edge pixel individually; no attempt is made to analyze the connectivity of the edge pixels or the local shape modeled by a set of connected edge pixels. In this regard, the potential of such detectors to capture the shape information of an object may be limited.

To explicitly characterize the local shapes of the object, Fergus *et. al* [11] extracted curved segments that have inflexion points at both it ends. Similarly, Jurie and Schmid [12] proposed a detector which searched over all positions and scales for sets of edges that describe salient circular arcs. Although both methods achieved good performance for object categorization, however these detectors imposed unnecessary restrictions on the class model to a homogenous structure composing only of *curves*. This inhibits its ability to learn complex class models.

2. PRIMITIVES EXTRACTION

We extract the primitives from the line edge map of the image. Our motivation for using the line edge map to extract the primitives stems from cognitive psychology studies [13, 14] which show that the humans are able to accurately identify an object from its line drawing. Given that the line drawing contains only the structural information of the object, this implies that the line edge map contains the structural information that is necessary to recognize an object.

2.1 Extraction of Quadrangles

We extract the quadrangles by grouping a pair of line segments of the line edge map. However, naïve pairing of the line segments leads to the construction of quadrangles, many of which do not provide meaningful structural description of the object. To this end, we represent the color image by its salient colors and exploit the lengths, closeness and orientations of the line segments and the perceptual uniformity of the salient colors within the quadrangle to evaluate the usefulness of the quadrangles. Our choice for these factors is motivated by the seminal psychological studies of Helson and Fehrer [15] who found that the human can more easily recognize and discriminate a uniformly colored rectangular shape than the other shapes. In addition, the pairing of long line segments that are close together explicitly models the Focal Point and the Proximity Laws of Gestalt Psychology [16]. By exploiting these factors, we identify those quadrangles that appeal to human visual attention

and hence may be useful for category-level object detection.

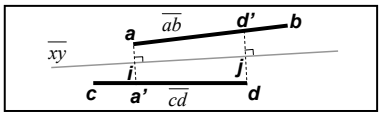


Figure 2: Pairing of $\overline{ab}$ and $\overline{cd}$ to define the quadrangle *aa'dd'*.

Fig. 2 shows two line segments $\overline{ab}$ and $\overline{cd}$. Line $\overline{xy}$ is the bisector of the angle formed by extending $\overline{ab}$ and $\overline{cd}$. If $\overline{ab}$ and $\overline{cd}$ are parallel, then line $\overline{xy}$ is defined as the line that lies between and is parallel to both $\overline{ab}$ and $\overline{cd}$. We denote the projection of points a and d onto $\overline{xy}$ as points i and j respectively. The overlap portion of $\overline{ab}$ and $\overline{cd}$ can then be represented as $\overline{ij}$. We denote the length of $\overline{ij}$ as l_{ij}. Let $\overline{ai}$ meet $\overline{cd}$ at point *a'*. Similarly, let $\overline{dj}$ meet $\overline{ab}$ at point *d'*. The quadrangle defined by $\overline{ab}$ and $\overline{cd}$ can then be represented as *aa'dd'*. The length and width of this quadrangle are given by l_{ij} and $\rho = \frac{l_{aa'}+l_{dd'}}{2}$ respectively. We denote $\overline{ij}$ as the central axis of the quadrangle. Let $1_A(x)$ define an indicator function in which x represent a pair of line segments and A denote the set of pairs of line segments that satisfy *Rule 1* : $\left(\frac{l_{ij}}{l_{ab}} \geqslant 0.5\right)$ *OR* $\left(\frac{l_{ij}}{l_{cd}} \geqslant 0.5\right)$. The spatial quality of the quadrangle defined by line segments $\overline{ab}$ and $\overline{cd}$ can then be calculated as

$$f\left(\overline{ab},\overline{cd}\right) = \frac{l_{ij}}{\rho} \times 1_A\left(\overline{ab},\overline{cd}\right) \times \cos\left(\theta\right) \quad (1)$$

where θ is the smaller intersecting angle between $\overline{ab}$ and $\overline{cd}$. If $\overline{ab}$ is parallel to $\overline{cd}$, θ is zero and according to the psychological studies of [15], such rectangular areas will capture the human visual attention more readily. We highlight that *Rule 1* is necessary to ensure that only line segments which have sufficient overlap will be considered to define a quadrangle. In our work, we choose a decision threshold of 0.5 to decide if two line segments have sufficient overlap. We observe that given two quadrangles, the term $\frac{l_{ij}}{\rho}$ in eq. (1) identifies the quadrangle that is composed of long line segments with close proximity to have better spatial quality as opposed to the quadrangle which is composed of short line segments that are far apart. This models the Focal Point Law of the Gestalt Psychology which suggests that given two line segments of different lengths, the human observer will inadvertently focus more on the longer line segment as opposed to the shorter line segment. Equally importantly, this term models the Proximity Law of the Gestalt Theory [17] which suggests that the two line segments that are close together will be perceived by the human observer to be one collective unit. Given that the indicator function and the $\cos(\theta)$ terms in eq. (1) model the human's attention towards a rectangular form, this formulation thus allow us to identify those quadrangles that are readily perceived by the human observer.

We retain those quadrangles whose spatial qualities are greater than zero. Let f_{max} and f_{min} denote respectively the maximum and minimum spatial qualities of the quadrangles whose spatial qualities are greater than zero. Eq. (1) can be normalized as follows,

$$\Phi\left(\overline{ab},\overline{cd}\right) = \frac{f\left(\overline{ab},\overline{cd}\right) - f_{min}}{f_{max} - f_{min}} \quad (2)$$

We now present our measurement which tightly integrates the number of pixels representing each color value, the perceptual distances between these color values and the color imbalance of the quadrangle to jointly evaluate the perceptual variations of the colors in a quadrangle. We first apply the approach in [18] to represent the image by its salient colors. Let m denote the number of salient colors within the quadrangle. For each salient color c_i in the quadrangle, we find its Euclidean distance $d(c_i,c_j)$ in the *Lab* color space to the other salient colors in the quadrangle and weight this distance by $w = \frac{\min(|\{c_i\}|,|\{c_j\}|)}{u}$, where u is the number of pixels in the quadrangle. The weighted distance represents the perceptual distance between c_i and c_j taking into account the population of each color. The maximum of the weighted distances are found for each c_i to calculate its contribution to the overall color variations of the quadrangle. These maxima are added to yield a measure of the color variations in a quadrangle defined by line segments $\overline{ab}$ and $\overline{cd}$,

$$g\left(\overline{ab},\overline{cd}\right) = \sum_{i=1}^{m} \max_j \left(w \times d\left(c_i,c_j\right)\right) \quad (3)$$

We retain those quadrangles whose color variations found by eq. (3) to be below λ. The threshold λ is determined as follows. We generate a set of five hundred rectangular patches of aspect ratio of 0.5. Each rectangular patch contains at most ten colors which are assigned to the pixels randomly. Using eq. (3), we calculate the values for the color variations of the rectangular patches and use these values to sort the rectangular patches. Starting with the rectangular patch of the least color variations, we present these patches to a human observer who decides if a patch has large perceptual variations in its colors. We record the value of the color variations λ_i of the first rectangular patch which is identified by the human observer as having large perceptual variations in its colors. This process is repeated three times, each time with a new set of rectangular patches. The average value of λ_i is then assigned to be λ. Let g_{max} and g_{min} denote respectively the maximum and minimum color variations among the quadrangles whose color variations found by eq. (3) to be below λ. The normalized color uniformity for the quadrangle defined by line segments $\overline{ab}$ and $\overline{cd}$ can be calculated as

$$\Psi\left(\overline{ab},\overline{cd}\right) = \frac{g_{max} - g\left(\overline{ab},\overline{cd}\right)}{g_{max} - g_{min}} \quad (4)$$

A measure of how much a quadrangle appeal to the attention of the human observer in terms of its spatial and color qualities can then be calculated as shown in eq. (5),

$$Appeal\left(\overline{ab},\overline{cd}\right) = 0.5 \times \Phi\left(\overline{ab},\overline{cd}\right) + 0.5 \times \Psi\left(\overline{ab},\overline{cd}\right) \quad (5)$$

In our work, we retain the top 85% of the quadrangles to construct the structural descriptors.

2.2 Extraction of Ellipses

In this section, we present our approach to locate and model the curved segments of the line edge map by ellipses. The novelty of our approach is that we continually pool the local information of the edge pixels together to achieve higher level understanding of the shapes of the curved segments. In addition, the parameters for the derived ellipses are continually refined using a close loop system driven by Gestalt Psychology [16]. Consequently, we are able to detect and model the curved segments with good visual perception.

We show the block diagram of our approach in Fig. 3. The *Partitions Extraction* component localizes the portions of the line edge map that can be modeled with ellipses. Specifically, we break the line edge map into a set of partitions such that each partition may correspond to the arc of at most one ellipse. We denote the i^{th} partition as $P_i = \bigcup_k l_k$, where l_k denote the k^{th} line segment within P_i. Given that the arc of an elliptical model is represented by a smooth and convex curve, a partition therefore cannot contain any corner or inflexion points. In light of this, we analyze the pair-wise geometrical relationships of the line segments within the partition to identify the corner and inflexion points of the line edge map and

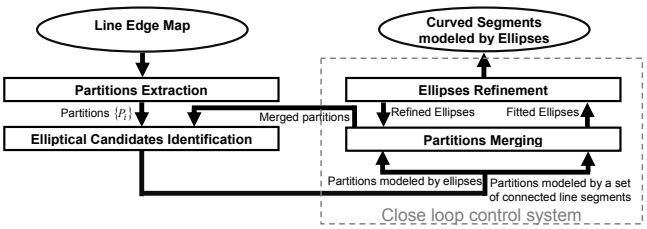


Figure 3: Block diagram of approach to model the curved segments by ellipses.

break the line edge map at these points. This ensures that the set of line segments within each partition will trace a smooth curve which turns either in the clockwise or anticlockwise direction. We adopt the method proposed by Chi and Leung [19] to detect the corner and inflexion points and denote the set of extracted partitions as $\{P_i\}$.

The *Elliptical Candidates Identification* component analyzes each partition P_i to determine how well its shape corresponds to an elliptical model. For this purpose, we fit an ellipse to the edge pixels represented in each partition and calculate the fraction of these edge pixels that overlap with the boundary of the fitted ellipse. Owning to the many advantages of [20], we adopt this method to fit the ellipse. Let E_i denote the ellipse that has been fitted to partition P_i and $\mathbb{B}(E_i)$ represent the set of boundary pixels of E_i. We define the set of edge pixels fitted by [21] to the line segments within P_i as $edge(P_i)$. In this case, P_i is identified to be an elliptical candidate if the quality of fit $f(E_i,P_i)$ as calculated in eq. (6) is greater than γ. In our work, we assign a low value of 0.2 to γ.

$$f(E_i,P_i) = \frac{|\mathbb{B}(E_i) \bigcap edge(P_i)|}{|edge(P_i)|} \tag{6}$$

The *Partitions Merging* component of the close loop system merges partitions according to Gestalt Psychology [16] and evaluates the quality of such merging by fitting an ellipse to the merged partition. Since the Gestalt Psychology considers how the human mind groups individual elements into a collective whole, therefore by exploiting the laws of Gestalt Psychology, we can identify the partitions that are likely to be perceived by the human observer to belong to a single elliptical unit. The merging procedure is modeled on the following laws of Gestalt Psychology:

- Law of Focal Point. This law states that given a visual representation, the mind will be pulled towards a point of emphasis. This focal point catches the viewer's attention and persuades the viewer to follow the visual message. In our work, we use the sum of the lengths of the line segments within a partition to determine if it should be given more emphasis.
- Law of Proximity. This law states that we tend to group elements that are spatially close together. To model this law, we consider two partitions for merging only if a line segment in one partition is connected to a line segment in the other partition.
- Law of Closure. This law states that our mind add missing elements to complete a pattern. Here the pattern is the elliptical curve and the missing elements are the additional partitions that better complete the curve. We model this law by accepting the merge of two partitions if the quality of fit of the merged partition is greater than that of either partitions.
- Law of Similarity. This law states that the mind will group elements that have some similarity to compose an entity. Here, we model this law pertaining to the *form* of the elements: Given two categories of partitions, the first containing partitions that are modeled by the boundary of an ellipse and the second containing partitions that are represented by a set of connected line segments, we merge those partitions that belong to the same category first before considering the partitions that belong to different categories.

The *Ellipse Refinement* component of the close loop system sieves out those line segments that do not have support from the other line segments within the partition in being modeled by an ellipse. Consequently, by removing such line segments, the remaining line segments of the partition will form a tightly integrated unit which collectively defines the elliptical portion of the line edge map. Let the set of edge pixels fitted by [21] to line segment l_m be denoted as $edge(l_m)$. We denote the a^{th} edge pixel of $edge(l_m)$ as $l_m(a)$. Let ellipse E_k denote the ellipse which is fitted to P_k and $E_k(j)$ denote the j^{th} boundary pixel of E_k. The *Ellipses Refinement* component calculates $Quality(E_k,l_m)$ for each line segment l_m of partition P_k and identifies those line segments whose $Quality(E_k,l_m)$ are below ϕ to be poorly modeled by E_k,

$$Quality(E_k,l_m) = \frac{\sum_{a=1}^{|edge(l_m)|} h(l_m(a),E_k)}{|edge(l_m)|} \tag{7}$$

where

$$h(l_m(a),E_k) = \begin{cases} 1 \text{ if } \exists i, E_k(i) \in \mathbb{B}(E_k) \text{ and} \\ \quad \langle E_k(i), l_m(a) \rangle \leq \varphi \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

and $\langle . \rangle$ denote the Euclidean distance. Let $\overline{P_k}$ denote the partition containing the line segments that have been identified to be poorly modeled by ellipse E_k. In this case, partition $P_k - \overline{P_k}$ contain those line segments that can be collectively modeled by an ellipse. We fit an ellipse explicitly to the edge pixels represented in partition $P_k - \overline{P_k}$ and denote the refined ellipse as $E_{k-\bar{k}}$. We feedback the set of line segments within partition $\overline{P_k}$ and the newly fitted ellipse $E_{k-\bar{k}}$ into the *Partitions Merging* component to determine if they can be merged with other partitions. The merging and feedback processes continue until no partitions can be merged. In our work, we assign ϕ and φ to have the values 0.5 and 2 respectively. As such, we can be assured that each line segment of partition $P_k - \overline{P_k}$ will be modeled by the boundary of ellipse $E_{k-\bar{k}}$ such that at least 50% of its representative edge pixels lie within two pixels from the boundary of ellipse $E_{k-\bar{k}}$. Consequently, accurate modeling of the curved segments by the ellipses is achieved. In our work, we use eq. (6) to calculate the quality of an ellipse and use only those ellipses whose qualities exceed 0.85 (which is considered high enough) to construct the descriptors.

3. STRUCTURAL DESCRIPTORS

We exploit the geometrical relationships and the spatial configurations of the neighboring primitives around each extracted ellipse to construct the descriptors. For the training image, we first segment the model from the image before extracting the primitives. In contrast, we do not perform any segmentation of the test image prior to extracting the primitives. In addition, for the training image, we randomly select between four to seven neighboring primitives around each ellipse to construct the descriptor. Correspondingly, for the test image, we consider all other primitives around each ellipse as its neighbors. In this aspect, a descriptor of the training image describes the *specific* structure around an ellipse while that of the teat image describes the *generalize* structure around an ellipse.

3.1 Extraction of geometrical relationships

We first present our method to capture the geometrical relationships between an ellipse and its neighboring primitives before showing how the structural descriptor is composed. We term the ellipse under consideration as an anchoring ellipse E_κ and denote its center as M_{E_κ} and the half lengths of its major and minor axes as α_κ and β_κ respectively. In addition, we denote the eccentricity of E_κ as ε_κ. We define the length and width of a neighboring quadrangle Φ as l_Φ and w_Φ respectively. We use a four dimensional vector

$\left[\frac{\alpha_\kappa}{\alpha_a}\ \frac{\beta_\kappa}{\beta_a}\ \frac{d(M_{E_\kappa},M_{E_a})}{\alpha_\kappa}\ \varepsilon_\kappa\right]^T$ to represent the geometrical relationships between E_κ and a neighboring ellipse E_a. Similarly, the geometrical relationships between E_κ and a neighboring quadrangle Φ are represented by the five dimensional vector $\left[\frac{\alpha_\kappa}{l_\Phi}\ \frac{\beta_\kappa}{w_\Phi}\ \frac{d(M_{E_\kappa},M_\Phi)}{\alpha_\kappa}\ \frac{l_\Phi}{w_\Phi}\ \varepsilon_\kappa\right]^T$. We highlight that the attributes in these vectors are invariant to translation, rotation and scaling.

3.2 Extraction of spatial relationships

We next discuss how the spatial relationships between an anchoring ellipse and its neighboring primitives are extracted. Corresponding, by integrating the spatial relationships with the geometrical relationships, we show our approach to construct the descriptor.

Fig. 4 shows an anchoring ellipse as the bold outline. We divide the region of space surrounding and within the ellipse into sixteen sub-regions and label them as S_A to S_P. These sub-regions represent sixteen angular sectors of three radial rings spaced exponentially in the half-lengths of the major and minor axes of the anchoring ellipse. Each sub-region delineates a space where the neighboring primitives, with respect to the ellipse, can be located. Given this, the structural descriptor consists of sixteen bins, in which each bin records the types and geometrical relationships of the neighboring primitives located in the corresponding sub-region with respect to the anchoring ellipse. In this aspect, the bins capture the spatial configurations of the primitives with respect to the anchoring ellipse.

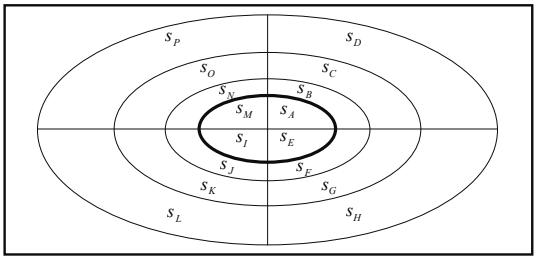


Figure 4: Diagram of sub-regions used to construct the descriptor.

3.3 Matching of structural descriptors

We match the descriptors as follows: We determine if there exists a circular shift in the descriptor of the test image such that Bin_j of the descriptor of the training image can be paired with Bin_j of the circular shifted descriptor of the test image, $1 \leq j \leq 16$. $Bin_{training}$ of the training image is considered to be paired with Bin_{test} of the test image if the correspondences for all geometrical attributes in $Bin_{training}$ can be found in Bin_{test}. Let $a_{training}$ and a_{test} be a geometrical attribute in $Bin_{training}$ and Bin_{test} respectively, in which both attributes are defined by the same type of primitives. We consider $a_{training}$ to correspond with a_{test} if $\frac{\min(a_{training},a_{test},k)}{\max(a_{training},a_{test},k)} \geqslant 0.5$, where k is a small number added to cater for situations when an attribute has value zero.

We highlight that since the primitives model the components of the object, the geometrical relationships of these primitives reveal the geometrical relationships of the components modeled by these primitives. Similarly, the spatial configurations of the neighboring primitives around an anchoring ellipse yield the description of its local structure. In this aspect, by exploiting the geometrical and spatial relationships between the primitives of the specified object in the training images, we learn descriptors that express the structural constraints of the object. Furthermore, as highlighted in the beginning of this section, a descriptor of the training image models the *specific* structure around a primitive while that of the test image models the *generalize* structure around a primitive. Corresponding, by matching the structural descriptors of the training images to that of the test image, we identify the descriptors from the test image that have a part of its structure similar to the model. Consequently, descriptors of the test image that have numerous matches suggest that these descriptors have many parts similar to the model. The likelihood that a descriptor of the test image will have many parts similar to the object model by accident is far lower than any initial match is in error. Such descriptors detect and localize the specified objects in the test image.

4. EXPERIMENTAL RESULTS

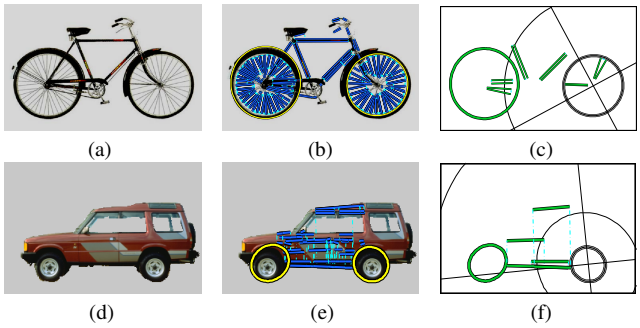


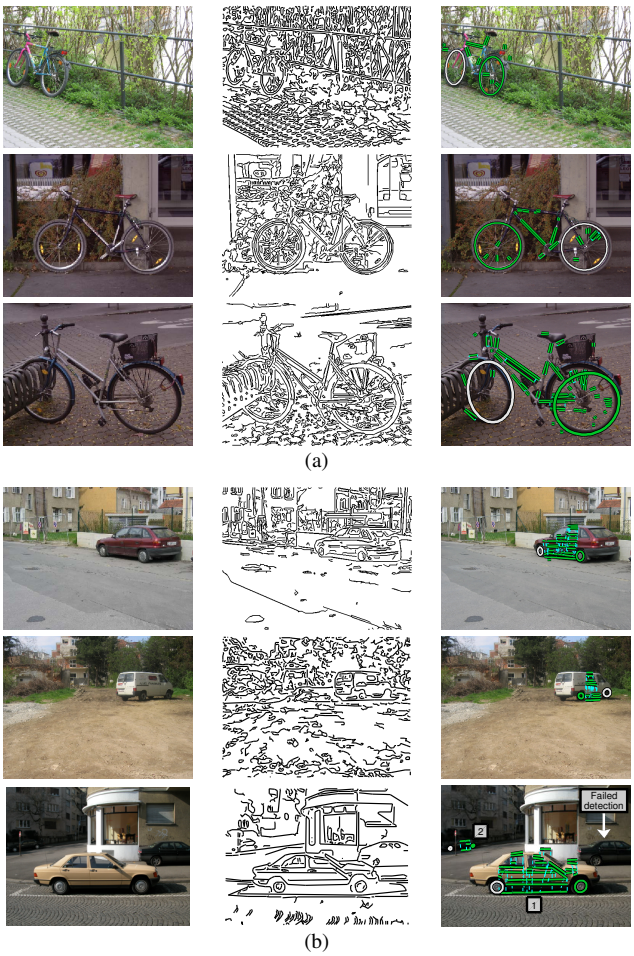
Figure 5: First column: Sample training images. Second column: Extracted primitives. Third column: Examples of structural descriptors.

We apply the structural descriptors composed by the ellipses and quadrangles to the detection of two object categories: Bicycles and cars. We learn each object model with a few training images. We show the examples of the training images in the first column of Fig. 5. For the bicycle images, we utilize the ellipses and quadrangles that model the wheels and frame of the bicycle to learn the descriptors. Specifically, we use an ellipse that model a wheel of the bicycle as the anchoring primitive. The structural descriptors are then constructed by selecting the ellipse that model the other wheel of the bicycle and six random quadrangles that model the frame of the bicycle as the neighboring primitives. As an example, we show in Fig. 5b, the ellipses (yellow outlines) and the quadrangles (blue outlines) extracted from the sample bicycle training image that are used to learn the descriptors. An example of the descriptor constructed from these primitives is shown in Fig. 5c where the anchoring ellipse is shown in gray and the neighboring primitives are shown in green. In the same spirit, we learn structural descriptors for the car model by exploiting the ellipse that model a side wheel of the car as the anchoring primitive. We select the ellipse that model the other side wheel of the car and three random quadrangles that model the side panel of the car as the neighboring primitives. We show the ellipses and quadrangles extracted from the sample training image of Fig. 5d that are used to learn the descriptors by the yellow and blue outlines respectively in Fig. 5e. One example of the descriptors constructed from these primitives is shown in Fig. 5f, where the anchoring ellipse is shown in gray and its neighboring primitives are shown in green.

We show the test images, its line edge maps and the detection results for each category in Fig. 6. We highlight that the objects in each category exhibit variations in scale and viewpoint and are of varying forms. In addition, as observed from the line edge maps, these objects are located in a cluttered background. For each test image, we identify the descriptor that have the most number of matches with the descriptors of the training images and shows its anchoring ellipse as the gray outline in the third column of Fig. 6a and 6b. In addition, we show the neighboring primitives of the anchoring ellipse which find matches with the descriptors of the training images by the green outlines. As observed, the most matched descriptors achieved good detection and localization of the objects.

Of particular interest is the car detection results shows in the third row of Fig. 6b. This test image shows three cars in the left, center and the right side of the image. We highlight that there are

significant scale variations between the three cars. In addition, there is substantial illumination variation in the scene, most notably in the left portion of the image for which the car present in this portion can hardly be visually detected. We label the top two matched descriptors as 1 and 2 respectively in the bottom row of Fig. 6b. As observed, the top matched descriptor correctly model the wheels and side panel of the car situated in the center of the image. More interestingly, the side wheels and panel of the car located in the left side of the image have been identified by the second most matched descriptor. Given that this car is located in the shadowed portion of the image and is notable small with respect to the size of the image, the detection of this car by the matched descriptor demonstrates the robustness of our approach. However, no descriptor pertaining to the car on the right side of the image has been detected. This is because we exploit the two side wheels and panel configurations of the car in the training images to learn the descriptors (see training image in Fig. 5d). Since a side wheel of this car is occluded, therefore no descriptor pertaining to this car was matched. The failure to detect this car is therefore due to the limited descriptors learned from the training images. In light of this, training methods e.g. [19] can be incorporated in our work to automatically learn a larger set of structural descriptors to represent an object model. With a larger and more informative set of descriptors, we can detect the other structures of the objects. Such detections can provide clues to the location of the object in the test image.



(a) (b)

Figure 6: (a) Bicycles detection. (b) Cars detection.

5. CONCLUSIONS AND FUTURE WORK

We have presented a new class of descriptors that exhibit the ability to yield meaningful structural description of the objects. We composed these descriptors from the quadrangles and ellipses that are extracted from the image. Promising category-level object detection results have been demonstrated. Future research will focus on the use of specialized training methods, such as that proposed in [19], to automatically learn the descriptors to represent an object. We point out here that although the most matched descriptor correctly localize the objects in the test images, there are nevertheless some descriptors which find false (albeit few) matches with the descriptors of the training images. These false matches can be filtered in future through higher level processing. Specifically, we can exploit the Hough transform to identify the descriptors of the test image that have a consistent interpretation of the object pose by using each matched descriptor of the training images to vote on all object poses. Finally, we highlight that although we have exploited the salient colors of the image to identify the useful quadrangles, we have not totally exhausted the use of the color information. Specifically, the colors/textures within the quadrangles and ellipses can be used to complement the current descriptors. Such representation provides a more discriminative description of the object and may be useful for specific object recognition.

REFERENCES

[1] Chuang, J., Ahuja, N., Lin, C., Tsai, C., Chen, C.: A potential-based generalized cylinder representationstar. Computers and Graphics (2004) 907918

[2] Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. ECCV (2000) 702718

[3] Rohr, K.: Towards model-based recognition of human movements in image sequence. CVIU (1994) 94–115

[4] Nalwa, V..: Line-drawing interpretation: Straight lines and conic sections. Trans. on PAMI (1988) 514–529

[5] Borel, R. In: A mathematical pattern recognition technique based on contour shape properties. Res. Foundation, Ohio State Univ. (1965)

[6] Fritzsche, D. In: Systematic method for character recognition. Res. Foundation, Ohio State Univ. (1961)

[7] Blum, H.: Biological shape and visual science. Theoretical Biology (1973) 205287

[8] Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. Trans. on PAMI (2002) 509–522

[9] Carmichael, O., Hebert, M.: Object recognition by a cascade of edge probes. BMVC (2002) 103–112

[10] Mikolajczyka, K., Zisserman, A., Schmid, C.: Shape recognition with edge-based features. BMVC (2003) 779–788

[11] R. Fergus, P.P., Zisserman, A.: A visual category filter for google images. ECCV (2004) 242–256

[12] Jurie, F., Schmid, C.: Scale invariant shape features for recognition of object categories. CVPR (2004) 90–96

[13] Irving, B., Ginny, J.: Surface versus edge-based determinants of visual recognition. Cognitive Psychology (1988) 38–64

[14] Bruce, V., Hanna, E., Dench, N., Healey, P., Burton, M.: The importance of mass in the line-drawings of faces. Applied Cognitive Psychology (1992) 619–628

[15] Helson, H., Fehrer, E.: The role of form in perception. The American Journal of Psychology (1932) 79–102

[16] Koffka, K. In: Principles of Gestalt Psychology. New York: Harcourt, Brace and Company (1935)

[17] Fisher, M., Gratto, K.S.: Gestalt theory: A foundation for instructional screen design. Journal of Educational Technology Systems (1998) 361–371

[18] Mojsilovic, A., Hu, J., Soljanin, E.: Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis. Trans. on IP (2002) 1238–1248

[19] Chi, Y., Leung, M.K.H.: Part-based object retrieval in cluttered environment. Trans. on PAMI (2007) 890–895

[20] Fitzgibbon, A., Pilu, M., Fisher, R.: Direct least squares fitting of ellipses. Trans. on PAMI (1999) 476–480

[21] Leung, M., Yang, Y.: Dynamic two-strip algorithm in curve fitting. Pattern Recognition (1990) 69–79