SHAPELEARNER: TOWARDS SHAPE-BASED VISUAL KNOWLEDGE HARVESTING

Zheng Wang[a] , Ti Liang[a]

[a]School of Computer Science and Technology, Shandong University - tjnkwz@gmail.com, liangti@foxmail.com

Commission III, WG III/5

KEY WORDS: Shape Knowledge Harvesting, Shape Matching, Shape Segmentation, Shape Synthesis

ABSTRACT:

The explosion of images on the Web has led to a number of efforts to organize images semantically and compile collections of visual knowledge. While there has been enormous progress on categorizing entire images or bounding boxes, only few studies have targeted fine-grained image understanding at the level of specific shape contours. For example, given an image of a cat, we would like a system to not merely recognize the existence of a cat, but also to distinguish between the cat's legs, head, tail, and so on. In this paper, we present ShapeLearner, a system that acquires such visual knowledge about object shapes and their parts. ShapeLearner jointly learns this knowledge from sets of segmented images. The space of label and segmentation hypotheses is pruned and then evaluated using Integer Linear Programming. ShapeLearner places the resulting knowledge in a semantic taxonomy based on WordNet and is able to exploit this hierarchy in order to analyze new kinds of objects that it has not observed before. We conduct experiments using a variety of shape classes from several representative categories and demonstrate the accuracy and robustness of our method.

1. INTRODUCTION

Motivation Over the last decade, we have observed an explosion in the number of images that are uploaded to the Internet. Sharing platforms such as Flickr and Facebook have long been driving forces in turning previously undistributed digital images into an abundant resource with tens of billions of images. This vast amount of data holds great potential to revolutionize the way computers organize and understand images. Deng (Deng et al., 2009) introduced ImageNet, a hierarchical organization of raw images, which has enabled major advances in object recognition to the point of computers outperforming humans in certain respects in object recognition (Russakovsky et al., 2015).

Still, current object recognition systems mostly operate at the coarse-grained level of entire images or of rectangular image bounding boxes, while segmentation algorithms tend to consider abstract distinctions such as between foreground and background.

In this work, we consider the next level of image understanding, aiming at a more fine-grained understanding of images by automatically identifying specific shape contours and the parts of objects that they portray. Analysis of objects with respect to their parts draws from cognitive research of the human vision systems. Shapes of parts play an important role in the lower stages of object recognition (Marr, 1976). Given a relatively small object part, humans can recognize the object when the part is sufficiently unique (Binford, 1971, Biederman, 1987). Such finer-grained image understanding has remained an open problem in computing, as it requires considerable background knowledge about the objects.

Contribution We introduce *ShapeLearner*, a system that learns the shapes of families of objects, together with their parts and their geometric realization, making the following contributions.

1. ShapeLearner requires only a small number of manually annotated seed shapes for bootstrapping and then progressively learns from new images. The core operation consists of a joint shape classification, segmentation, and annotation procedure. To solve this challenging central task, ShapeLearner automatically transfers visual knowledge of seen shapes to unseen images, accounting for both geometric and semantic similarity.
2. ShapeLearner can automatically analyse entirely new kinds of shapes, relying on an inference mechanism based on soft constraints, such as discrepancies between shape families, part uniqueness, etc. (see Figure 1). Once such a new shape has been classified, segmented, and annotated, the newly acquired knowledge is incorporated into ShapeLearner to further enhance its knowledge.
3. Rather than learning mere enumerations, the system acquires hierarchical knowledge about these parts, which is semantically more informative (Palmer, 1977, Hoffman and Richards, 1983). Additionally, different object categories are organized hierarchically as well, following the WordNet taxonomy (Fellbaum, 1998), e.g. to account for the relationship between a cup and a glass. This hierarchical organization is critical when jointly analyzing families of objects, due to the high degree of geometric variability of shapes at different levels of granularity.

2. RELATED WORK

2.1 Image Knowledge Harvesting

In recent years, several new methods have appeared to organize the growing amount of images on the Web. The most prominent of these is ImageNet (Deng et al., 2009), a hierarchically organized image knowledge base intended to serve as the visual counterpart of WordNet (Fellbaum, 1998). While ImageNet merely provides image-level labels, subsequent research aimed at localizing individual objects within those images using bounding boxes (Guillaumin and Ferrari, 2012). In our work, we focus on the specific shape contours of objects and analyse their subparts.

A semantic infrastructure was introduced by AIM@SHAPE (Falcidieno et al., 2004) to provide a semantic representation of shapes on the Internet. For a survey on content based 3D shape retrieval please refer to (Tangelder and Veltkamp, 2008). In this context, textual taxonomies have been also utilized to constrain interactive tools and generate consistent segmentations and annotations of 3D shapes (Robbiano et al., 2007) and images (Russell et al., 2008). Hierarchical taxonomies have been also used to train classifiers

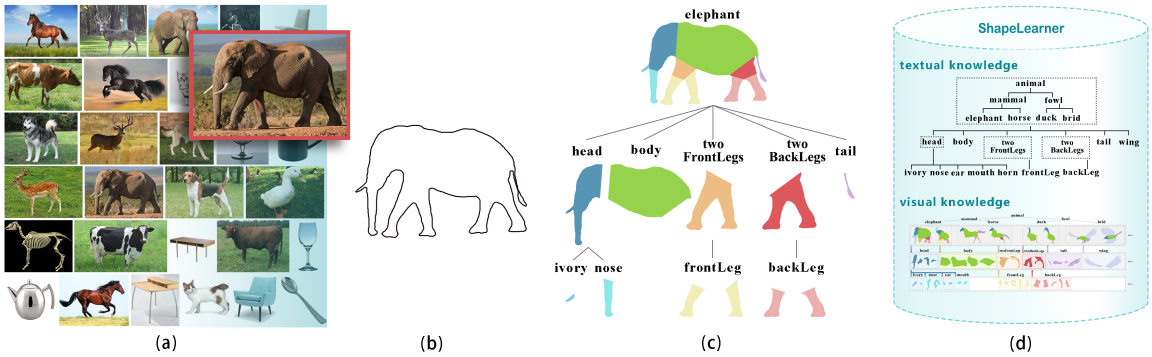


Figure 1: The proliferation of images on the Web (a) enables us to extract shapes to train ShapeLearner (b), a 2D shape learning system that acquires knowledge of shape families, geometrical instances of their inner parts and their inter-relations. Given an unknown shape (c), the system automatically determines a classification, segmentation, and hierarchical part annotation (d).

and constrain organization of videos from the Internet (Song et al., 2010). In (Patterson and Hays, 2012, Patterson et al., 2014) a crowd-sourced taxonomy was introduced for the organization and classification of a large-scale scene database.

With the arrival of low-cost RGB-D sensors, there has been a growing demand for classification and organization of large object families represented by depth images. Lai et al. (Lai et al., 2011) present a hierarchical RGB-D knowledge base of 51 classes organized according to WordNet taxonomy. Similarly, synthetic 3D CAD models have been organized according to WordNet in a hierarchical base denoted 3DNet (Wohlkinger et al., 2012).

2.2 Segmentations and Semantic Relationships

Zhang (Zhang et al., 2011b) observe that semantic relations of parts may be typically expressed by a common structure which is shared among objects in a class. Hence, they learn a set of classifiers for the relation of verb-object in a class of images. Similarly, graph structures have been introduced for representing semantic relations of parts which are learned from image sets (Malisiewicz and Efros, 2009, Chen et al., 2012). Recently Zhang (Zhang et al., 2014) organized co-occurrences and contextual relationships of images in a graph which assists annotation and retrieval of Web-scale images. Nevertheless, these methods focus on processing general images and scenes while we believe to be the first to focus on individual shape classes, their inner parts and geometries.

Grammars have been suggested to represent the visual information in images as high-level generative models (Girshick et al., 2011, Zhang et al., 2011a). Grammar-like descriptors for visual words and visual phrases may be defined to enhance image processing and recognition (Zhang et al., 2011a). Recently, Chen (Chen et al., 2013) learn object relationships in images from their probabilistic structural patterns and geometrical characteristics. Their work shares a common goal with us of leveraging semantic object relations for the construction a large-scale visual knowledge-base. Nevertheless, our analysis focuses on semantic relations at the sub-part level, while theirs is at the global scene level and object-object relations.

Multiple instances of objects and parts within a class, provide important contextual information which is utilized for joint learning and segmentation (Rother et al., 2006, Chum and Zisserman, 2007, Batra et al., 2010, Vicente et al., 2011, Chai et al., 2011, Kang et al., 2011). For example, a consistent segmentation of similar 2D objects may be achieved from multiple segmentations which are interconnected with a constrained graph (Kim et al., 2012) . Similar to us, Huang (Huang et al., 2014) recently presented a data-driven approach for simultaneous segmentation and annotation of free-hand sketches. They utilize a database of 3D objects and parts which are superimposed with the 2D sketch to infer the best fitting structure. In contrast to us, their input shapes also contain interior information which provides important hints in the recognition process, especially for ambiguous shapes.

3. OVERVIEW AND KNOWLEDGE MODEL

3.1 High-Level Perspective

ShapeLearner constructs a relational hierarchy that indexes 2D shapes by utilizing taxonomic knowledge of object shape classes and their inner parts. Our goal is to progressively acquire such knowledge by transferring information about indexed shapes onto new ones.

We manually index shapes in several categories as seeds (e.g., mammals, fowls, home appliances). This involves segmenting images collected via Google Images to separate the objects from their environment. Objects are then manually segmented further into meaningful parts and labeled following the WordNet taxonomy. ShapeLearner captures this information about parts and their relations in a tree-like hierarchy by connecting parts to their siblings and ancestors. This can be seen as a knowledge base with *is-a*, *has-a*, *has-part*, and *has-shape* relationships.

ShapeLearner includes a knowledge transfer algorithm for understanding unknown shapes. ShapeLeaner accounts for both shape geometry and high-level semantical relations from its previously acquired knowledge to infer the correct classification and segmentation of the new object shape. This is illustrated in Figure 2: Given an unknown shape, we compute a raw set of segmentation candidates considering merely the shape's geometry. We determine additional candidates by matching with geometrically similar shapes and transferring their segmentation. This yields a set of segmentation hypotheses about the unknown shape. ShapeLearner then transfers its knowledge onto the shape by relying on an inference step to remove false hypotheses and select a valid segmentation that complies with the shape's hierarchical taxonomy. Finally, ShapeLearner transfers this knowledge back by indexing the new shape and progressively updating its store of visual knowledge.

3.2 ShapeLearner's Knowledge

ShapeLearner is directly linked to the WordNet (Fellbaum, 1998) taxonomy, which provides a hierarchical semantic organization of classes.

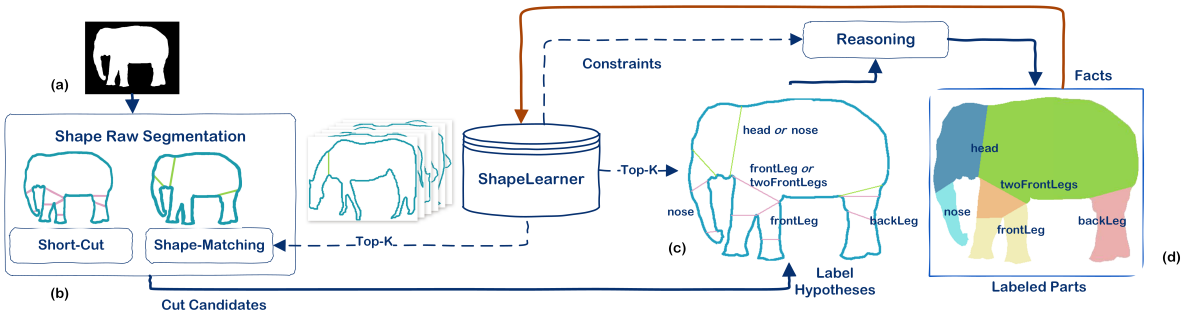


Figure 2: Workflow diagram. Given an unknown 2D shape (a), ShapeLearner first determines segmentation candidates by leveraging short cut and shape matching information (b). The system uses its acquired knowledge to label candidates (c). Finally, it makes use of reasoning to prune false hypotheses and infer a classification and semantic segmentation of the shape (d).

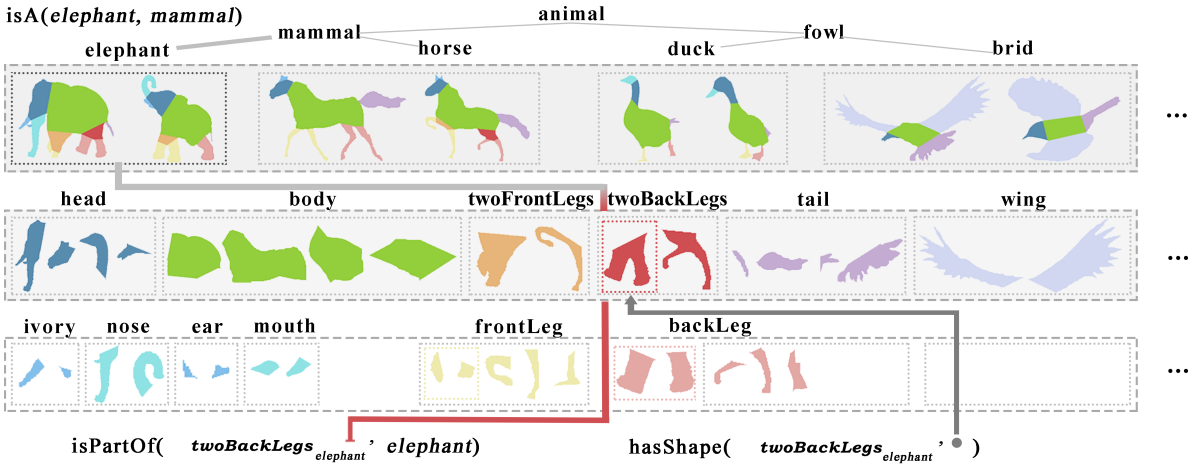


Figure 3: A snapshot of the knowledge in ShapeLearner's hierarchy, zooming in on mammals and fowls. We show also a subset of the relational facts *isA*, *isPartOf* and *hasShape*.

Focusing on a subset of this taxonomy, we take the `isA` class and additionally harvest knowledge for `isPartOf` and `hasShape` facts (e.g. *isPartOf* `(Leg, Human)`, *hasShape* `(Baseball, Round)`) Thus, ShapeLearner acquires knowledge of an object's *shape*, its *parts*, and *shape of the parts* (see Figure 3).

We begin by defining the four basic concepts that ShapeLearner relies on:

- Shapes $\mathcal{S} = \{s_0, s_1, ..., s_{n_\mathcal{S}}\}$ define the contour of independent 2D objects in an image.
- Classes $\mathcal{C} = \{c_0, c_1, ..., c_{n_\mathcal{C}}\}$ define a category (e.g., species) of objects in the data base.
- Parts $\mathcal{P} = \{p_0, p_1, ..., p_{n_\mathcal{P}}\}$ define a decomposition of a shape into meaningful components.
- Labels $\mathcal{L} = \{l_0, l_1, ..., l_{n_\mathcal{L}}\}$ define the textual annotations for each part.

Initially, a seed set of parts is manually preprocessed and transferred into ShapeLearner. In this step, the user manually annotates parts in shapes with labels from WordNet (e.g. *head*, *tail*, etc.) as well as semantic relations (e.g. *hasShape(elephant,elephantShape), isA(elephant,mammal), isPartOf(tail,elephant)*). ShapeLearner stores this information in a hierarchical structure (see Figure 3).

Next, we use ShapeLearner to infer the following knowledge in a statistical manner:

- *Part_number*: the number of parts per class may be fixed or bounded (e.g. a horse has 2 front legs, an elephant has 1 trunk).
- *Part_distinctiveness*: Shape classes may have discriminate parts defined by the frequency of a part in all classes (e.g. the *elephant* class has trunks as a distinct part within the class of *mammals*). Part_distinctiveness is at the core of shape classification and disambiguation. The part distinctiveness score for a part p in class $c \in \mathcal{C}$ is calculated as the inverse fraction of classes containing this part: $\frac{|\mathcal{C}|}{|p \in c|} \geq \epsilon, \epsilon = |\mathcal{C}|$.

4. SHAPE ANALYSIS

Classification and semantic segmentation of an unknown object shape typically pose a chicken-egg problem: we may require information about the one in order to solve the other. Given an unknown 2D shape, ShapeLearner jointly solves for both classification and semantic segmentation by relying on an inference procedure to reason from its knowledge in accordance with statistical constraints and the shape geometry. In fact, it jointly optimizes classification, segmentation, as well as part annotation. We next provide the technical details of this process.

4.1 Shape Segmentation Hypotheses

Given an unknown shape of an object, we compute a set of possible part candidates specified by different cuts in the shape (see cuts in Figure 5(c)). Initially, we compute cuts accounting merely for the shape geometry, applying the short-cut rule of (Luo et al., 2014), which is motivated by the human vision system. This method yields somewhat consistent cuts tracking the geometric features of the shape contour. Nevertheless, our algorithm does not require an exact segmentation into meaningful parts but only

a loose approximation. A somewhat reasonable segmentation is sufficient at this step.

Next, ShapeLearner transfers additional segment hypotheses from its existing knowledge to further enrich the candidate set. Shape matching plays an important role in adding new cuts that further enrich segmentation and compensate when the short-cut geometry-based method is insufficient. For instance, in Figure 5(a), the smooth elephant head could not be segmented by the short-cut method.

To accomplish this, ShapeLearner finds the best matching shapes in its existing collection and transfers their segmentation onto the input shape. Shape matching is performed using the inner-distance similarity metric (Ling and Jacobs, 2007). We found this method suitable as it is computationally efficient, rotation-invariant, and robust with respect to other state-of-the-art 2D contour matching techniques (e.g., (Belongie et al., 2001)).

Following the inner distance metric (Ling and Jacobs, 2007), we define $C(\pi(A,B))$ as the matching cost value for two shapes A and B. In a nutshell, given two shapes A and B, described by their contour point sequences $p_1, p_2 ... p_n$ and $q_1, q_2, ... q_m$, respectively, we use the χ^2 statistic to compare points histograms similarity presented the cost value of $c(p_i, q_j)$. We compute the optimal matching between A and B, denoted as $\pi : (p_i, q_{\pi(i)})$, using dynamic programming. We define the minimum cost value by $C(\pi) = \sum_{i=1}^{n} c(i, \pi(i))$ and the number of matching points is $M(\pi) = \sum_{i=1}^{n} \delta(i)$, where $\delta(i) = 1$ if $\pi(i) \neq 0$, $\delta(i) = 0$ if $\pi(i) = 0$.

Next, we define a cut, i.e. $cut_A(p_i, p_j)$, as the 2D line connecting contour points p_i,p_j in shape A. Thus, to transfer $cut_A(p_i, p_j)$ from shape A in ShapeLearner onto the input shape B, we simply use the computed shape matching π and transfer $cut_A(p_i, p_j)$ to $cut_B(q_{\pi(i)}, q_{\pi(j)})$ (Figure 5(b)).

To reduce noise in the segmentation candidates, ShapeLearner considers only the top $k_1 = 3$ best matching shapes in its collection. Additionally, it relies on the following constraints to remove noisy cuts (Figure 4):

- cut should be located in the interior of the shape.
- when cuts intersect each other, only the one corresponding to the longest contour is kept.
- if two cuts are too close together, specifically $\|cut_B(d) - cut_B(e)\|_2 \leq \epsilon$, where $\epsilon = 0.01 \times |shape\_points|$, they are merged together.

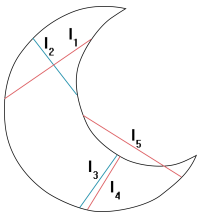


Figure 4: Cut constraints remove all red cuts.

4.2 Shape-Class and Part-Label Hypotheses

At this point, ShapeLearner has an unknown shape and a set of unlabeled segments, so the shape may belong to different classes and a cut may have different labels. Thus, ShapeLearner next annotates segments with possible label hypotheses from its knowledge and computes a valid segmentation that conforms to its acquired knowledge, by cleaning false segments and label hypotheses.

We assign a unique ID for each cut in the shape and denote an hypothesis as the pair *(cut,label)[.]*. Additionally, we define class hypotheses as *(shape,class)[.]*. A hypothesis may become a fact *(cut,label)[1]* or be evaluated as false *(cut,label)[0]*, following an inference process (e.g. label(cut@9, nose)[1], class(shape@1, elephant)[0]). Note that each cut corresponds to a part, so label(cut@9, nose)[1] equals label(part@9, nose)[1]. ShapeLearner matches the input shape against its knowledge and select the top $k = 5$ best matching shapes using the inner distance metric. This yields multiple class and label assignments for the hypotheses.

We define the cut confidence weight with respect to the top k resulting set as follows. Given a cut c_j, label l_i, and hypotheses: label(cut@j, l_i)[.], the confidence weight of cut c_j with label l_i is calculated as $w_{c_j,l_i} = \alpha \times p_1 + (1-\alpha) \times p_2$, ($\alpha = 0.6$ in our experiments), based on two factors:

- p_1: the confidence of assigning label l_i to cut c_j is $\frac{h_l}{k}$, where h_l is the frequency of label l_i in the top k result set.
- p_2: if a cut has many possible label hypotheses (say $l_1, l_i, \ldots, l_m$), the confidence for each part is defined by the part shape matching $w'_{c_j,l_i} = M_{l_i}(\pi)/C_{l_i}(\pi)$. Then $p_2 = \frac{w'_{c_j,l_i}}{\sum_l w'_{c_j,l}}$.

Similarly, we define the class confidence weight with respect to the top k result set as follows. Given the unknown part_shape s_j, class c_i and hypothesis class(shape@j, c_i)[.], the confidence of class c_i with respect to the top k result set is calculated as $w_{s_j,c_i} = \frac{h_c}{k}$, where h_c is the number of hits for class c_i.

4.3 Shape Inference

ShapeLearner jointly solves for a consistent classification and labeling by pruning noisy hypotheses and searching for the optimum class and labels assignment with respect to its knowledge constraints. We formulate this problem as an Integer Linear Programming (ILP) that considers both cut labels and shape classes to yield a consistent set of truth value hypotheses.

We formulate the ILP variables as follows:

- $x_{p,l} \in \{0,1\}$ denotes label(part, label) hypothesis $l \in \mathcal{L}$ for part $p \in \mathcal{P}$.
- $y_{s,c} \in \{0,1\}$ denotes class(shape, class) hypothesis $c \in \mathcal{C}$ for shape $s \in \mathcal{S}$.

For each shape s, the objective function maximizes the overall confidence of hypotheses (where $w_{x_{p,l}}$ and $w_{y_{s,c}}$ are the confidence weights for cut and class hypotheses respectively, $w_{x_{p,l}} = w_{x_{c,l}}$ in the previous step):

$$\max \sum_{p \in \mathcal{P}, l \in \mathcal{L}} w_{x_{p,l}} \cdot x_{p,l} + \sum_{c \in \mathcal{C}} w_{y_{s,c}} \cdot y_{s,c}$$

subject to the following constraints derived statistically from the knowledge collection.

4.4 Class Constraints.

- A shape s can be assigned to one class at most:

$$\sum_{c \in \mathcal{C}} y_{s,c} \leq 1$$

- A shape class assignment should conform to its distinctive parts (if any). Denoting $(l,c) \in DPC$ as the pair set (distinctive part, class), then:

$$\forall p \in \mathcal{P} \wedge c \in \mathcal{C} \wedge (l,c) \in DPC, x_{p,l} - y_{s,c} \leq 0$$

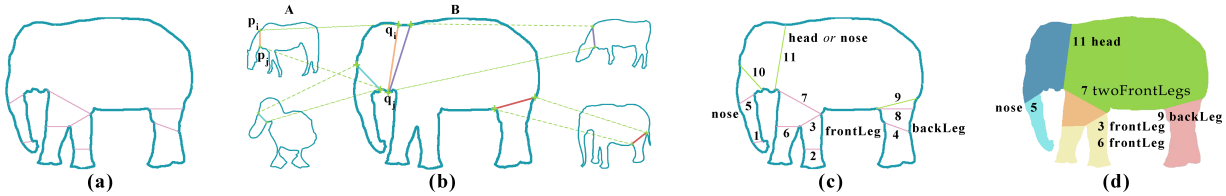


Figure 5: Semantic segmentation of an elephant. Given an unsegmented shape, cuts are computed from the geometry (a), as well as transferred from similar shapes (b). This yields multiple class hypotheses (c) which are prunen, yielding a correct semantic segmentation and annotation of the shape (d).

- A shape class should not consist of parts that do not belong to the class (according to `isPartOf`)

  $\forall p \in \mathcal{P} \wedge c \in \mathcal{C} \wedge (l,c) \notin isPartOf(p,c), x_{p,l} + y_{s,c} \leq 1$

4.5 Label Constraints.

- Part inclusion should conform to ShapeLearner's part hierarchy. Denoting H_{P} as the set of inclusion part pairs (i.e. iff $(l,l') \in H_{\mathrm{P}}$, then l' includes l and $isPartOf(p,p')$), and $\subset$ refers to "included by" then:

  $\forall p,p' \in \mathcal{P}, l,l' \in \mathcal{L} \wedge (l,l') \in H_{\mathrm{P}} \wedge p \not\subset p', x_{p,l} + x_{p',l'} \leq 1$

  and

  $\forall p,p' \in \mathcal{P}, l,l' \in \mathcal{L} \wedge (l,l') \notin H_{\mathrm{P}} \wedge p \subset p', x_{p,l} + x_{p',l'} \leq 1$

- The number of parts in a shape class should conform to the class Part_number. Denoting the number of parts as n_{P}, then:

  $$\sum_{p \in \mathcal{P}} x_{p,l} \leq n_{\mathrm{P}} c, l$$

  Note that we require the number of parts to be less than or equal to n_{P} due to possible occlusions of the shape in the image (c.f. the back leg in Figure 5).

After reasoning, the cleaned facts (i.e., label(part, label)[1] and class(shape, class)[1]) are integrated into ShapeLearner's knowledge base. The shape of each part is added as `hasShape(part, 'part-shape')`. Given a shape of a new class not yet in ShapeLearner, parts of the new class are identified by knowledge transfer. If the new class name is `X`, new facts are added as `isPartOf(part, X)` and `hasShape(part, part-shape)`.

5. RESULTS

We now present a thorough set of experiments to evaluate ShapeLearner.

5.1 Labeling Accuracy

To quantify ShapeLearner's output quality, we rely on a pixel-based metric to evaluate the parts segmentation (Huang et al., 2014). Given a segmented part, we measure its overlap with the ground-truth part as the number of pixels that are correctly labeled in the overlap vs. the incorrect ones. A part is considered adequately labeled if a reasonable percentage (precision $> 75\%$) of pixels are in the overlap. The terminology is as follows.

- **True Positive (TP):** correct cut/pixel label
- **True Negative (TN):** correct removed cut/pixel label
- **False Positive (FP):** a cut/pixel label supposed to be removed but not removed
- **False Negative (FN):** a cut/pixel supposed to be labeled, but removed.

Given these, we can use the standard definition of precision as $\frac{TP}{TP+FP}$, recall as $\frac{TP}{TP+FN}$, and $F_1 = \frac{2TP}{2TP+FP+FN}$.

5.2 Baselines

Given all part hypotheses, we evaluate our method against two simpler baselines:

- **N:** the inference includes Part_number constraints.
- **N+D** the influences includes Part_number and Part_distinctiveness constraints.

Table 1 provides an evaluation of the segmentation and classification for these baselines with respect to precision, recall, and the F_1 measure. Our method outperforms these baselines in most cases (except re call in some cases). Figure 6 provides examples of a summary of this a subset of this evaluation, illustrating F_1 results of baselines and our method.

In Figure 7(a), we demonstrate the scalability of our method with respect to the number of initial seeds for classes with size larger than 50. Note that precision, recall, and F_1 of the segmentation increase as the number of seeds gets larger. After 20 seeds, ShapeLearner converges and the improvement becomes marginal. Therefore, 20 seeds appear to be a reasonable threshold in our experiments. This shows that a small number of seeds can represent a shape-space well and adding more seeds can be redundant. Figures 7(b) and 7(c) show graphic comparisons between baselines and our inference mechanism according to the values in Table 1. We see that even for a small number of seeds, our method outperforms other baselines and has very good precision, recall, and F_1. Nevertheless, segmentation of the deer and the cow classes is challenging and stayed below the average due to ambiguities (see Table 1). Small parts, such as ears and horns, when represented only as contours are not sufficiently distinct (even for the human eye).

Our classification (Table 1, bottom part) also outperforms the baselines on average. Nevertheless, for some classes we did not improve over baselines since their contour was relatively similar with no distinctive parts. For example, the small horn of the deer is similar to the ear of the horse. As horns are distinctive for deers, a horse shape will be classified as a deer if its ear is labeled as a horn. Ribs in the skeleton may also be ambiguous since they are similar to legs in size and orientation. Similarly, the cat's tail may be recognized as a back leg in unique situations when the tail hangs down and the cat's back legs are occluded.

Similarly for skeletons, classification did not provide excellent results due to the large similarity between skeletons and their

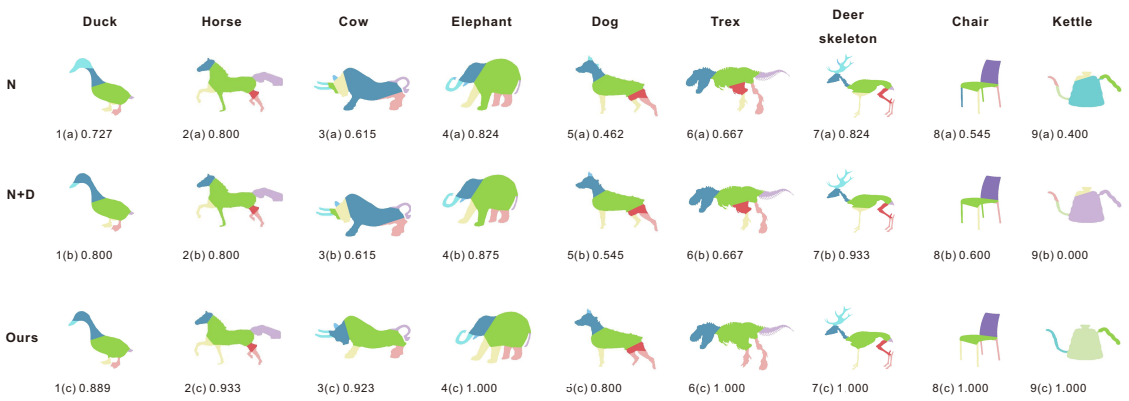


Figure 6: Representative results by our method and baseline solutions. The F1 measure is shown below each result.

counterpart (living) mammals. Furthermore, the variety of shapes in the skeleton class is not high, which does not provide distinct and meaningful recognition knowledge and constraints. Nevertheless, the segmentation of the skeleton was successful for the same reasons of similarity with mammals and transferring their knowledge back.

The experiments about knowledge transfer in Table 3 show that even without seeds from the same class, seeds of classes from the same category could help for transferring segmentation and annotation. When the seeds from different classes are more similar with the test shape, the performance is even better than the direct segmentation and annotation (see tiger and bedroom lamp in Table 3). Morphological differences between tortoise and other reptiles are quite big. Thus the performance of transfer segmentation is not very good, compared with other reptiles.

5.3 Evaluation and Comparison

Although our paper has a different aim, we compare our method with a segmentation algorithm (Huang et al., 2014) aiming at analysing hand-drawn sketches. The main difference is that their method is aimed at analysing the brush strokes, which may contain significant information on the shape's interior (interior sketches), while ours considers only the contour. From their dataset, we select all object classes with contours and compare the average segmentation and annotation precision (see Table 4). While their algorithm can resolve many ambiguities due to occlusions based on the interior brush strokes, our method nevertheless gives superior results on a majority of classes, demonstrating the effective power of ShapeLearner's knowledge. For the airplane and vase classes, our method was inferior due to the large variety (airplanes) and non-distinctiveness of parts (vases). Unfortunately, we could not perform a more in-depth comparison (e.g. w.r.t. occlusions and a larger variety of classes) since their code is not publicly available.

Our method can infer a semantically correct segmentation even for classes that are not currently indexed in ShapeLearner. In Figure 8, three shapes of new classes (a lion, rhinoceros, and camel), were properly segmented and annotated (including, for example, the rhinoceros' horn) by ShapeLearner without having been given prior knowledge about these classes.

6. CONCLUSION

We have introduced a novel system that organizes 2D shapes in a hierarchical structure and learns to process new images and even

	Huang	Ours		Huang	Ours
airplane	66.2%	65.8%	lamp	89.3%	94.9%
candelabra	56.7%	68.5%	rifle	62.2%	67.2%
fourleg	67.2%	80.9%	vase	63.1%	51.0%
human	64.0%	94.1%			

Table 4: Comparision with Huang et al. [2014] in precision.

(a) Lion (b) Ostrich (c) Alpaca

Figure 8: Semantic segmentation of three new shapes (without prior indexing of these classes by ShapeLearner).

new shape categories. Our system starts with a seed set of annotated shapes but then augments its knowledge by automatically processing new images and shapes. We derive a set of statistical constraints that we apply to correctly classify and segment an unknown input shape. We transfer hypotheses based on visual similarity, which we then validate using an integer linear programming reasoning method. Our experiments show that, after seeding, ShapeLearner is able to collect valuable knowledge about shapes from uncategorized images.We additionally present several applications as use-cases of ShapeLearner, showcasing enhanced shape processing and manipulation.

In future work, we would like to extend ShapeLearner to focus not only on 2D shapes represented by their contours, but also to analyse the interior textures. While a reduction to 2D shape contours reduces some of the noise, it results in a minimalist geometric representation.We are currently exploring deep learning methods to analyse interior textures.By going beyond it, ShapeLearner could thus also be extended to handle object shapes with severe shape occlusions.

REFERENCES

Batra, D., Kowdle, A., Parikh, D., Luo, J. and Chen, T., 2010. icoseg: Interactive co-segmentation with intelligent scribble guidance. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec., pp. 3169–3176.

System	Mammals	Home Appliances	Misc. Artifacts	Foods	Reptiles	Fowls	Skeletons	All Avg.	
N	68.3%	86.0%	88.5%	100.0%	74.5%	65.8%	57.3%	77.2%	***Prec.***
N+D	69.2%	90.4%	92.3%	100.0%	74.5%	66.6%	59.8%	79.0%	
Ours	79.4%	92.5%	92.6%	100.0%	84.1%	75.6%	71.8%	85.1%	
N	85.5%	93.3%	93.7%	100.0%	85.4%	90.4%	76.4%	89.2%	***Recall***
N+D	85.1%	92.4%	94.0%	100.0%	85.4%	90.3%	77.1%	89.2%	
Ours	86.9%	93.6%	94.4%	100.0%	83.0%	91.4%	80.8%	90.0%	
N	74.8%	88.1%	90.2%	100.0%	78.9%	74.3%	64.5%	81.6%	***F1***
N+D	75.3%	90.9%	93.0%	100.0%	78.9%	74.9%	66.4%	82.8%	
Ours	82.2%	92.5%	93.3%	100.0%	82.9%	81.1%	74.6%	86.7%	
N	65.3%	91.4%	87.8%	93.3%	92.5%	87.8%	61.9%	82.9%	***Class***
N+D	72.0%	93.2%	92.6%	93.3%	95.0%	88.9%	58.8%	84.8%	
Ours	71.9%	93.7%	92.6%	93.3%	95.0%	89.3%	59.3%	85.0%	

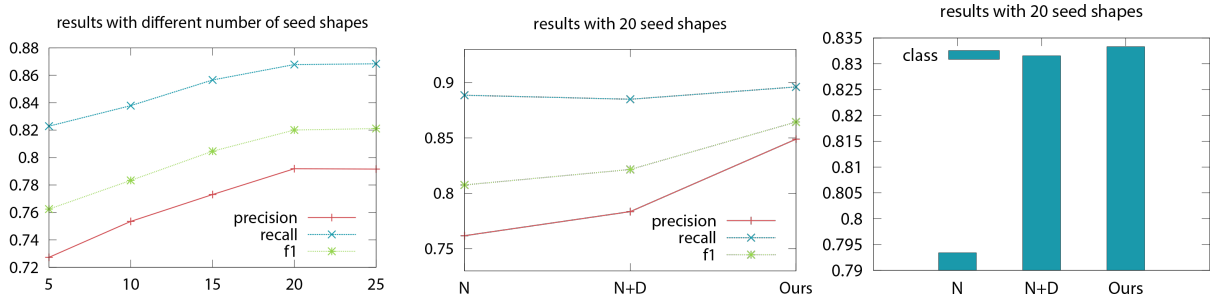
Table 1: Experimental results for segmentation and annotation (top) and classification (bottom).

System	Mammals					Home Appliances			Misc. Artifacts		Foods	Reptiles		Fowls		Skeletons		
	Elephant	Cow	Deer	Horse	Cat	Vase	Hairdryer	Broom	Rifle	Axe	Mushroom	Tortoise	Crocodile	Duck	Bird	Mammals	Dinosaur	
N	74.6%	62.4%	80.6%	64.5%	63.3%	67.8%	96.7%	96.7%	59.1%	93.3%	100.0%	65.6%	68.8%	63.2%	67.4%	62.2%	52.5%	
N+D	75.8%	63.2%	81.7%	64.6%	65.5%	73.3%	96.7%	96.7%	78.2%	93.3%	100.0%	65.6%	68.8%	63.8%	69.2%	64.3%	55.3%	***Prec.***
Ours	86.0%	71.4%	87.0%	77.9%	79.6%	73.3%	96.7%	96.7%	79.9%	93.3%	100.0%	78.2%	81.4%	74.4%	79.2%	75.1%	68.4%	
N	90.5%	81.3%	87.7%	83.9%	80.3%	80.0%	96.7%	96.7%	85.3%	93.3%	100.0%	80.9%	84.8%	89.5%	90.6%	78.4%	74.5%	
N+D	88.9%	79.6%	87.7%	83.1%	80.6%	78.3%	96.7%	96.7%	86.8%	93.3%	100.0%	80.9%	84.8%	87.5%	92.2%	78.4%	75.9%	***Recall***
Ours	91.1%	84.2%	90.0%	86.2%	84.8%	78.3%	96.7%	96.7%	88.5%	93.3%	100.0%	81.1%	89.9%	86.7%	93.1%	82.8%	78.8%	
N	81.2%	69.7%	83.0%	72.0%	70.1%	72.1%	96.7%	96.7%	67.8%	93.3%	100.0%	71.4%	75.4%	72.6%	75.6%	68.3%	60.6%	
N+D	81.3%	69.5%	83.7%	71.8%	71.5%	75.0%	96.7%	96.7%	81.6%	93.3%	100.0%	71.4%	75.4%	72.4%	77.5%	69.6%	63.1%	***F1***
Ours	88.0%	76.5%	87.8%	81.2%	81.4%	75.0%	96.7%	96.7%	83.3%	93.3%	100.0%	79.2%	84.9%	78.6%	84.1%	77.9%	71.4%	
N	94.4%	53.1%	90.4%	37.8%	86.0%	90.0%	96.7%	76.7%	69.0%	70.0%	93.3%	96.6%	90.0%	83.5%	83.3%	34.0%	89.9%	
N+D	94.4%	60.5%	80.9%	76.7%	76.0%	86.7%	100.0%	76.7%	93.1%	70.0%	93.3%	96.6%	100.0%	89.9%	76.7%	37.7%	79.8%	***Class***
Ours	94.4%	59.3%	80.9%	75.6%	76.0%	86.7%	100.0%	76.7%	93.1%	70.0%	93.3%	96.6%	100.0%	91.1%	76.7%	38.9%	79.8%	

Table 2: Excerpts for segmentation and annotation (top) and classification (bottom).

Method	Feline			Reptiles				Lamp			Canine			
	Cat	Leopard	Tiger	Tortoise	Crocodile	Lizard	Gecko	Desk Lamp	Floor Lamp	Bedroom Lamp	Dog	Wolf	Fox	
Precision	79.6%	73.7%	75.1%	78.2%	81.4%	88.2%	88.7%	92.6%	94.6%	85.0%	79.8%	71.5%	77.3%	***Direct***
Recall	84.8%	91.1%	78.1%	81.1%	89.9%	78.5%	82.4%	90.7%	96.4%	85.6%	92%	84.6%	84.9%	
F1	81.4%	80.3%	76.2%	79.2%	84.9%	82.6%	84.9%	91.4%	95.2%	83.2%	84.5%	76.1%	80.2%	
Precision	66.7%	70.1%	86.7%	46.2%	71.3%	78.4%	81.2%	88.9%	92.9%	91.7%	78.0%	69.8%	74.3%	***Trans.***
Recall	60.0%	72.6%	70.2%	52.9%	69.5%	71.9%	77.5%	87.0%	100.0%	78.9%	69.2%	77.1%	66.1%	
F1	62.0%	69.6%	76.7%	48.6%	69.8%	74.8%	78.3%	87.7%	95.2%	82.1%	71.6%	71.2%	68.6%	

Table 3: Experimental results for with seeds (top) and with only transfer (bottom).



(a) Scalability. (b) Comparison between baselines. (c) Classification results.

Figure 7: Experimental results graphs. In (a) we show the scalability of the average precision, recall and F1, and in (b) the comparison with other baselines. In (c) we show classification precision comparison with other baselines.

Belongie, S., Malik, J. and Puzicha, J., 2001. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, pp. 509–522.

Biederman, I., 1987. Recognition-by-components: A theory of human image understanding. Psychological Review 94, pp. 115–147.

Binford, T. O., 1971. Visual perception by computer. In: Proc. of the IEEE Conf. on Systems and Control (Miami, FL).

Chai, Y., Lempitsky, V. S. and Zisserman, A., 2011. Bicos: A bi-level co-segmentation method for image classification. In: Proc. Int. Conf. on Comp. Vis., pp. 2579–2586.

Chen, N., Zhou, Q.-Y. and Prasanna, V., 2012. Understanding web images by object relation network. In: Proceedings of the 21st International Conference on World Wide Web, pp. 291–300.

Chen, X., Shrivastava, A. and Gupta, A., 2013. NEIL: extracting visual knowledge from web data. In: Proc. Int. Conf. on Comp. Vis., pp. 1409–1416.

Chum, O. and Zisserman, A., 2007. An exemplar model for learning object classes. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.

Falcidieno, B., Spagnuolo, M., Alliez, P., Quak, E., Houstis, C. and Vavalis, E., 2004. Towards the semantics of digital shapes: the AIM@SHAPE approach. In: European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. Bradford Books.

Girshick, R., Felzenszwalb, P. and McAllester, D., 2011. Object detection with grammar models. In: Proceedings of Advances in Neural Information Processing Systems (NIPS).

Guillaumin, M. and Ferrari, V., 2012. Large-scale knowledge transfer for object localization in imagenet. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec., pp. 3202–3209.

Hoffman, D. and Richards, W., 1983. Parts of recognition. COGNITION 18, pp. 65–96.

Huang, Z., Fu, H. and Lau, R. W. H., 2014. Data-driven segmentation and labeling of freehand sketches. ACM Trans. on Graphics.

Kang, H., Hebert, M. and Kanade, T., 2011. Discovering object instances from scenes of daily living. In: Proc. Int. Conf. on Comp. Vis.

Kim, E., Li, H. and Huang, X., 2012. A hierarchical image clustering cosegmentation framework. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec., pp. 686–693.

Lai, K., Bo, L., Ren, X. and Fox, D., 2011. A large-scale hierarchical multi-view RGB-D object dataset. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1817–1824.

Ling, H. and Jacobs, D. W., 2007. Shape classification using the inner-distance. PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON.

Luo, L., Shen, C., Liu, X. and Zhang, C., 2014. A computational model of the short-cut rule for 2d shape decomposition. CoRR.

Malisiewicz, T. and Efros, A. A., 2009. Beyond categories: The visual memex model for reasoning about object relationships. In: NIPS.

Marr, D., 1976. Early processing of visual information. Philosophical Transactions of the Royal Society of London B: Biological Sciences 275(942), pp. 483–519.

Palmer, S. E., 1977. Hierarchical structure in perceptual representation. Cognitive Psychology pp. 441–474.

Patterson, G. and Hays, J., 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.

Patterson, G., Xu, C., Su, H. and Hays, J., 2014. The sun attribute database: Beyond categories for deeper scene understanding. International Journal of Computer Vision pp. 59–81.

Robbiano, F., Attene, M., Spagnuolo, M. and Falcidieno, B., 2007. Part-based annotation of virtual 3d shapes. In: Proceedings of the 2007 International Conference on Cyberworlds, pp. 427–436.

Rother, C., Minka, T., Blake, A. and Kolmogorov, V., 2006. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec., pp. 993–1000.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV).

Russell, B., Torralba, A., Murphy, K. and Freeman, W., 2008. Labelme: A database and web-based tool for image annotation. Interl. Journal of Computer Vision (1-3), pp. 157–173.

Song, Y., Zhao, M., Yagnik, J. and Wu, X., 2010. Taxonomic classification for web-based videos. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec., pp. 871–878.

Tangelder, J. W. and Veltkamp, R. C., 2008. A survey of content based 3d shape retrieval methods. Multimedia Tools Appl. 39(3), pp. 441–471.

Vicente, S., Rother, C. and Kolmogorov, V., 2011. Object cosegmentation. In: Proc. IEEE Conf. on Comp. Vis. and Pat. Rec., pp. 2217–2224.

Wohlkinger, W., Aldoma, A., Rusu, R. and Vincze, M., 2012. 3dnet: Large-scale object class recognition from cad models. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 5384–5391.

Zhang, S., Tian, Q., Hua, G., Huang, Q. and Gao, W., 2011a. Generating Descriptive Visual Words and Visual Phrases for Large-Scale Image Applications. IEEE Transactions on Image Processing pp. 2664–2677.

Zhang, S., Tian, Q., Hua, G., Huang, Q. and Gao, W., 2014. Objectpatchnet: Towards scalable and semantic image annotation and retrieval. Comput. Vis. Image Underst. pp. 16–29.

Zhang, X., Zha, Z.-J. and Xu, C., 2011b. Learning "verb-object" concepts for semantic image annotation. In: ACM Multimedia, pp. 1077–1080.