Visual Analogies: A Framework for Defining Aspect Categorization

P. Daphne Tsatsoulis, Bryan A. Plummer, and David Forsyth

University of Illinois at Urbana-Champaign
{tsatsou2,bplumme2,daf}@illinois.edu

Abstract. Analogies are common simple word problems (calf is to cow as x is to sheep?) and we use them to identify analogies between images. Let $\mathcal{I}[\mathcal{A}, \theta]$ be an image of object $\mathcal{A}$ at view θ. We show how to learn to choose an image $\mathcal{I}$ such that $\mathcal{I}[\mathcal{A}, \phi]$ is to $\mathcal{I}[\mathcal{A}, \theta]$ as $\mathcal{I}$ is to $\mathcal{I}[\mathcal{B}, \theta]$. We introduce a framework to identify an image of a familiar object at an unfamiliar angle and extend our method to treat unfamiliar objects. By doing so, we identify pairs of objects that are good at finding new views of one another. This yields an operational notion of aspectual equivalence: objects are equivalent if they can predict each other's appearance well.

1 Introduction

Objects look different when looked at from different directions, an effect known as aspect. In this paper we attack a problem with little history: how do we decide which objects share aspectual properties? We do this based on a predictive notion of aspect without using geometric or appearance information.

We propose the analogy task (visualized in Figure 1(i)) to model the relationship between different objects and aspects in a category-independent manner. Our experiments show we can transfer the knowledge from an analogy to recognize an object from an unseen aspect. By doing so we implicitly capture 3D structure through a learning-based approach without explicit models. We introduce the idea of Aspectual Categories, equivalence classes between objects that capture shared aspectual properties. We use analogies to define three problems:
(I) Aspect Transfer A system should take the views it has of an object and use them to identify the same object in new views. This is difficult since we cannot expect to have images of an object at all angles.
(II) Aspect Transfer across Objects A system should be able to take many views of one particular object and use them to predict the changes that occur when the viewpoint changes for images of a *different object*.
(III) Aspect Categorization We cannot expect a successful aspect transfer across all pairs of objects. For example, two views of a box are unlikely to make it easier to predict a second view of a hedgehog. We would like to know which pairs of objects support aspect transfer.

The results of our Aspect Transfer across Objects experiments are used to create Aspect Categories. These categories summarize which objects share as-

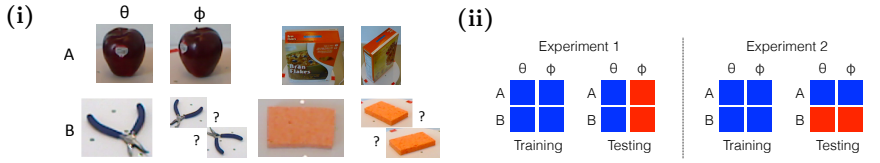

Fig. 1. (i) Given images of object A at angles θ and ϕ, with an image of object B at angle θ, we can correctly predict which image of object B completes the visual analogy. (ii) Two unique sets (red, blue) of angles (left) or objects (right) are defined for each experiment. The angle (ϕ) or object ($\mathcal{B}$) of the fourth element in the 4-tuple has not been seen in training.

pectual properties. Objects are equivalent if they correctly predict views of each other during the Aspect Transfer across Objects experiment.

Contributions: 1. Our method can identify a particular new view of a known object by analogy. **2.** This method can be used to decompose a set of objects into aspectual equivalence classes. Two objects are equivalent if one can use views of one object to predict views of the other. **3.** We evaluate our framework on the three problems presented and develop baselines to which future work can compare using the RGBD-3D Dataset.

2 Related Work

Aspect in object recognition: There are three main strategies for handling aspect. One is to build a **comprehensive representation of aspectual phenomena** (an **aspect graph**; review in [1], critique in [2], summary of results in [3]). This usually results in complex representations and has fallen into disuse.

Another, to represent an object using **aspect-enriched models**. In the extreme, rather than build a "car" recognizer, one might build "frontal-car", "lateral-car" and "overhead-car" recognizers. Usually, these multiple classes are compacted into a single model, assembled from local patches, tied together by observation [4], with geometric reasoning [5], with statistical reasoning [6, 7], or with a combination [8–10]. This strategy is expensive in data. However, one may interpolate missing aspects [11], or interpolate models corresponding to missing aspects [12].

Generally, there is little direct study of aspect transfer across objects. The models of Xiang and Savarese [9, 10] decompose objects into salient parts from different views, and record when each is visible. It is likely that numerous objects could be encoded by a single part decomposition of this form, but using distinct appearance models (for example, matchboxes and omnibuses). An alternative is to build **aspect invariant features** which are known only from distinct special constructions (e.g. [13] for specialized cases; ([14], [15]) for human activities; [16] for ASL). A disadvantage of this is it handles categories independently.

A number of papers using visual analogies have been published in the last year. Though similar in topic to our paper we still tackle a substantially novel problem. We differentiate between transferring across aspects and across objects unlike [17] or [18] along with providing corresponding experiments for those problems. Furthermore, unlike [18] we preform experiments on real data. Unlike [19] and [20] we formulate the problem as an analogy problem.

3 Task

Given two pairs (A, A') and (B, B'), an analogy exists if the relation between A and A' is equivalent to the one between B and B'. We apply this concept to the image domain and predict the equivalence.

We write objects as $\{\mathcal{A}, \mathcal{B}\}$, and view angles as $\{\theta, \phi\}$; write $\mathcal{I}([\mathcal{A}, \theta])$ for an image of object $\mathcal{A}$ at view angle θ. We can operationalize this analogical reasoning by choosing a function F that accepts four images such that,

$$F(\mathcal{I}[\mathcal{A}, \theta], \mathcal{I}[\mathcal{A}, \phi], \mathcal{I}[\mathcal{B}, \theta], \mathcal{I}[\mathcal{B}, \phi]) > F(\mathcal{I}[\mathcal{A}, \theta], \mathcal{I}[\mathcal{A}, \phi], \mathcal{I}[\mathcal{B}, \theta], \mathcal{J})$$

where $\mathcal{J}$ is any image other than $\mathcal{I}[\mathcal{B}, \phi]$. We require that this property be true for all $\mathcal{A}$, $\mathcal{B}$, θ, ϕ. Given F with this property, it is straightforward to identify a view of object $\mathcal{B}$ at view ϕ. We search for the image $\mathcal{I}$ such that $F(\mathcal{I}[\mathcal{A}, \theta], \mathcal{I}[\mathcal{A}, \phi], \mathcal{I}[\mathcal{B}, \theta], \mathcal{I})$ is largest. We conduct three experiments to answer the following questions:

I Aspect Transfer: Can our method generalize over angle? Can it correctly select $\mathcal{I}[\mathcal{B}, \phi]$ if trained with object $\mathcal{B}$ but not angle ϕ.

II Aspect transfer Across Objects: Can our method generalize over objects? Can it correctly select $\mathcal{I}[\mathcal{B}, \phi]$ if trained with angle ϕ but not object $\mathcal{B}$.

III Aspect Categorization: Which object-pairs are easy to generalize over?

3.1 Experimental Design

Experimental design matters a lot for this problem, particularly the test-train split. It is easy to confuse train and test data when creating 4-tuples by including a prediction image in a training 4-tuple and in a testing 4-tuple. It is also easy to make too-easy examples by setting $\theta = \phi$. Similarly, the task can also be made too difficult by using disjoint sets of objects and angles in test and train.

I: Transfer Across Aspect The method learns how an image can change from angle θ to ϕ_{train} and needs to predict if the change from θ to ϕ_{test} is correct. To do so, all angles are split into two sets, test and train, with 4 angles in each. These were used to create a train set $\{\mathcal{I}[\cdot, \theta], \mathcal{I}[\cdot, \phi_{\text{train}}], \mathcal{I}[\cdot, \theta], \mathcal{I}[\cdot, \phi_{\text{train}}]\}_{\text{train}}$ and a test set $\{\mathcal{I}[\cdot, \theta], \mathcal{I}[\cdot, \phi_{\text{test}}], \mathcal{I}[\cdot, \theta], \mathcal{I}[\cdot, \phi_{\text{test}}]\}_{\text{test}}$. The training set was used to define θ and ϕ_{train}. The testing set was used to define ϕ_{test}. Please see Figure 1(i) for a visualization.

II: Aspect Transfer Across Objects In this experiment, the method learns how an object $\mathcal{A}$ changes angles and needs to predict whether a new object $\mathcal{B}_{\text{test}}$ changed in the same way. To do so, we took all objects and split them into two sets. These created a training set $\{\mathcal{I}[\mathcal{A}, \cdot], \mathcal{I}[\mathcal{A}, \cdot], \mathcal{I}[\mathcal{B}_{\text{train}}, \cdot], \mathcal{I}[\mathcal{B}_{\text{train}}, \cdot]\}_{\text{train}}$ and

a testing set $\{\mathcal{I}[\mathcal{A},\cdot],\mathcal{I}[\mathcal{A},\cdot],\mathcal{I}[\mathcal{B}_{\text{test}},\cdot],\mathcal{I}[\mathcal{B}_{\text{test}},\cdot]\}_{\text{test}}$. The training set was used to define $\mathcal{A}$ and $\mathcal{B}_{\text{train}}$. The testing set was used to define $\mathcal{B}_{\text{test}}$.

Evaluation Metrics for Experiments I and II For experiments I and II we wanted to evaluate the method over many angle- (or object-) pairs. To do so, we used a pooled AUC. We regarded each tuple $\{\mathcal{I}[\mathcal{A},\theta],\mathcal{I}[\mathcal{A},\phi],\mathcal{I}[\mathcal{B},\theta],\mathcal{I}[\mathcal{B},\phi]\}$ as positive (whatever $\mathcal{A},\mathcal{B},\theta,\phi$) and all others as negative. We then computed the AUC. By doing so, we obtained a summary of performance for each case.

III: Aspect Categorization In this experiment we address question III posed above: Can our method generate categories with similar aspect? In this Experiment we use the results from Experiment II to create our aspect categories. We regard objects $\mathcal{A}$ and $\mathcal{B}$ as aspectually similar if, for many angle pairs θ,ϕ, our method accurately finds $\mathcal{I}([\mathcal{B},\phi])$. We can capture this notion by computing the AUC over angles for pairs of objects. The result is the AUC when using object $\mathcal{A}$ to predict object $\mathcal{B}$ for all pairs of objects.

We cluster on the AUC using an agglomerative clustering with complete-link distance. Objects that share a cluster are better at predicting one another.

4 Method

4.1 Gradient Tree Boost

We use Gradient Boosting [21] to predict which 4-tuples are analogies. Gradient boosting constructs an ensemble of learners: in our case, regression trees. Each iteration, 1... m, learns a new tree, h_m, from the residuals of the previous iteration's forest, $F_{m-1}(\mathbf{x})$. More specifically, it regresses the input features, $\mathbf{x}$, against the negative gradient of the loss function evaluated at the predicted output $f(\mathbf{x})$. It also learns a weight for each tree. The forest is grown for a number of iterations, m, and evaluated using a loss function and its gradient.

Exponential Loss The exponential loss, $\mathscr{L}(y,f(\mathbf{x}))=e^{-yf(\mathbf{x})}$, with derivative, $\frac{\partial}{\partial f(\mathbf{x})}=-ye^{-yf(\mathbf{x})}$, penalizes examples $\mathbf{x}$ for which $f(\mathbf{x})$ is incorrectly predicted.

AUC Loss The Area Under the Receiver-Operating-Characteristic (AUC) is a cost function that cannot be directly applied as a per-example loss function. However, because we are trying to find the best fit for a visual analogy, it makes sense to define a loss that relates prediction $f(\mathbf{x})$ to all other predictions in the way a ranker would. We want positive examples, f_i, to score higher than all negative examples, f_j, which is captured by the AUC. A high AUC reflects a method in which there are a low number of false positives and a high number of true positives. We modify the AUC cost function to define a loss as [22], $\mathscr{L}(\mathbf{y},f(\mathbf{x}))=1-\text{AUC}(\mathbf{y},f(\mathbf{x}))=1-\frac{1}{|S_+||S_-|}\sum_{i\in S_+}\sum_{j\in S_-}\mathbb{1}(f_i-f_j>0)$ Where S_+ are examples with a true positive label and S_- are examples with a true negative label. The indicator function $\mathbb{1}(f_i-f_j>0)$ is not differentiable and needs to be approximated with the sigmoid function, $\sigma(f_i-f_j)=\frac{1}{1+e^{-\beta(f_i-f_j)}}$. As $\beta\to\infty$ this approximates the indicator function's step-like behavior. The partial derivative of the loss function with respect to a single point is:

$$\frac{\partial}{\partial f_{a\in S_+}}=\frac{-\beta}{|S_+||S_-|}\sum_{j\in S_-}\frac{e^{\beta(f_a-f_j)}}{(1+e^{\beta(f_a-f_j)})^2}\,,\qquad \frac{\partial}{\partial f_{a\in S_-}}=\frac{\beta}{|S_+||S_-|}\sum_{i\in S_+}\frac{e^{\beta(f_i-f_a)}}{(1+e^{\beta(f_i-f_a)})^2}$$

4.2 Data and Features

We used the training set of the RGBD-3D Dataset [23] that contains 51 types of objects at 360 angles. We used the crops of the first example in each class and angles $\{0, 45, 90, 135, 180, 225, 270, 315\}$. We extracted features, $h(x_{a,t})$, for every item, a, at angle, t, using a Deep Residual Network described in [24]; a 152 layer network pre-trained on imagenet. We used the activations before the last fully connected layer (res5c). We used a combination of per-image features for a set of images: $f(x_{\mathcal{A},\theta}, x_{\mathcal{A},\phi}, x_{\mathcal{B},\theta}, x_{\mathcal{B},\psi}) = [\Delta h_{\mathcal{A},\{\theta,\phi\}}, \Delta h_{\{\mathcal{A},\mathcal{B}\},\theta}, \Delta h_{\mathcal{B},\{\theta,\phi\}}, \Delta h_{\{\mathcal{A},\mathcal{B}\},\phi}, \Delta h_{\mathcal{A},\{\theta,\phi\}} - \Delta h_{\mathcal{B},\{\theta,\phi\}}, \Delta h_{\{\mathcal{A},\mathcal{B}\},\theta} - \Delta h_{\{\mathcal{A},\mathcal{B}\},\phi}]$. Feature $f(x_{\mathcal{A},\theta}, x_{\mathcal{A},\phi}, x_{\mathcal{B},\theta}, x_{\mathcal{B},\psi})$ is a positive example, $+1$, if $\phi = \psi$ and a negative example, -1, if $\phi \neq \psi$.

4.3 The Model

We use Gradient Boosting [21] to model visual analogies. Each regression tree was grown using the entire training set. We ran the method for a maximum of 100 iterations (for a maximum of 100 trees per forest). Each tree was grown until a minimum leaf size of 10, 50, or 100 was reached. We used both Exponential and AUC loss methods and varied the AUC loss parameter β to be 100 or 1000.

5 Results

We evaluate predictions using the area under the ROC curve (AUC). This metric best captures the results in a biased dataset. There are almost an order-of-magnitude more negative than positive examples (for each correct angle there are seven incorrect angles). The AUC simulates a forced-choice test in which the system must pick between a positive and negative example. For the aspect transfer and aspect transfer across object experiments we report AUC where we computed the AUC over all examples. For the aspect category experiment we report a per-object-pair $\text{AUC}_{\text{obj-pair}}$ which was pooled over angles for each pairing of objects. This AUC captures how well two objects predict each other.

Loss	MLS = 10	50	100	MLS = 10	50	100
AUC, $\beta = 100$	0.6728	0.6925	**0.6950**	0.5676	0.5717	**0.5726**
$= 1000$	0.6563	0.6904	0.6910	0.5676	0.5717	**0.5726**
Exponential	0.5980	0.6534	0.6393	0.5543	0.5415	0.5595

Table 1. Results of the **angle-split** (left) and object-split (right) experiments for two losses and varying minimum leaf size (MLS). Chance performance is 0.5.

(I) Aspect Transfer: We identify $\mathcal{I}([\mathcal{B}, \phi])$ given $\{\mathcal{I}([\mathcal{A}, \theta]), \mathcal{I}([\mathcal{A}, \phi]), \mathcal{I}([\mathcal{B}, \theta])\}$ with no instance of angle ϕ in the training set. We correctly identify $\mathcal{I}([\mathcal{B}, \phi])$ with a pooled AUC of 0.6950 when using the AUC loss function with a β parameter equal to 100 and a minimum leaf size of 100. Using the exponential loss

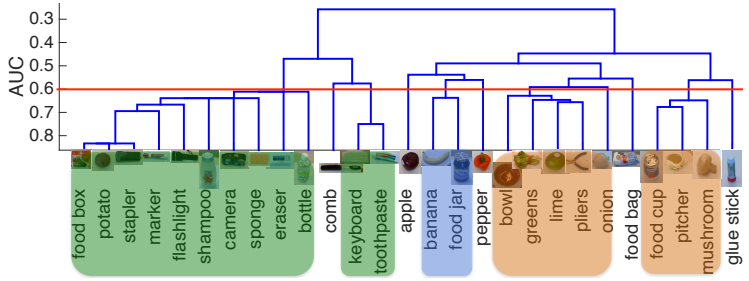

Fig. 2. The complete-clustering over two-object AUCs. Objects that quickly cluster together predict each other's rotation with high AUC. For a chosen AUC of 0.6 (red line) clusters are formed that depend on the shape of the objects. There are boxy clusters (green), round clusters (orange), and one cluster (blue) is a mix of shapes.

with the same minimum leaf size gave an AUC of 0.6393 suggesting gains were made by penalizing based on the AUC loss.

(II) Aspect Transfer Across Objects: We identify $\mathcal{I}([\mathcal{B}, \phi])$ given $\{\mathcal{I}([\mathcal{A}, \theta]), \mathcal{I}([\mathcal{A}, \phi]), \mathcal{I}([\mathcal{B}, \theta])\}$ with no instance of object $\mathcal{B}$ in the training set. We correctly identify $\mathcal{I}([\mathcal{B}, \phi])$ with a pooled AUC of 0.5726 when using the AUC loss with a β parameter equal to 100 and a minimum leaf size of 100.

Transfer across objects (II) is more difficult than the transfer across aspect (I) because the model only used comparisons between *other* object pairs when training. Predicting a rotation to an unseen viewpoint is not as extreme a task as predicting the viewpoint changes of unseen objects.

(III) Aspect Categorization: We clustered objects that had high AUC when used together in an analogy. Figure 2 provides is an illustration of groups of objects that best predict each other's orientations. Neighboring objects have better AUC performance when used to predict each other. For example, it would be better to compare a toothpaste with a keyboard than a water-bottle because the toothpaste-keyboard analogies had an AUC ≈ 0.75 and the toothpaste-bottle analogies had an AUC ≈ 0.52. The lower the bar that connects two objects (the higher the AUC) the better they are at predicting each other. The clustering shows how the method picks up on strong coordinated behavior between objects.

The clusters formed respect general geometric descriptions of the objects. The {food stapler, marker, sponge,...} and {keyboard, toothpaste} clusters are made of boxy objects that have clear differences between most of their orientations. Our features are able to pick up on strong coordinated behavior such as the boxy structure in the previous cluster and the curves in circular objects or objects labels in the {food-cup,pitcher, pitcher, mushroom} cluster. Since we have clustered based on the two-object AUC we can tell which observed object $\{\mathcal{A}\}$ will best predict object $\{\mathcal{B}\}$. For a tolerable predictive AUC we have defined the cluster of objects with which to make predictions.

References

1. Bowyer, K., Dyer, C.: Aspect graphs: an introduction and survey of recent results. **2** (1990) 315–328
2. Faugeras, O., Mundy, J., Ahuja, N., Dyer, C., Pentland, A., Jain, R., Ikeuchi, K., Bowyer, K.: Why aspect graphs are not (yet) practical for computer vision. CVGIP **55**(2) (March 1992) 212–218
3. Forsyth, D., Ponce, J.: Computer Vision: a modern approach. Prentice-Hall (2002)
4. Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Schiele, B., Gool, L.V.: Towards multi-view object class detection. In: CVPR. (2006) 1589–1596
5. Huang, C.Y., Camps, O., Kanungo, T.: Object recognition using appearance-based parts and relations. In: CVPR. (1997) 877–83
6. Kushal, A., Schmid, C., Ponce, J.: Flexible object models for category level 3d object recognition. In: CVPR. (2007)
7. Lazebnik, S., Schmid, C., Ponce, J.: Semi-local affine parts for object recognition. In: British Machine Vision Conference. (2004)
8. Savarese., S., Fei-Fei, L.: 3d generic object categorization, localization and pose estimation. In: ICCV. (2007) 1–8
9. Xiang, Y., Savarese, S.: Estimating the aspect layout of object categories. (2012)
10. Xiang, Y., Savarese, S.: Object detection by 3d aspectlets and occlusion reasoning. In: 4th International IEEE Workshop on 3D Representation and Recognition. (2013)
11. Chiu, H.P., Kaelbling, L.P., Lozano-Perez, T.: Virtual training for multi-view object class recognition. In: CVPR. (2007) 1–8
12. Savarese, S., Fei-Fei, L.: View synthesis for recognizing unseen poses of object classes. In: ECCV. (2008)
13. Forsyth, D., Mundy, J., Zisserman, A., Coelho, C., Heller, A., Rothwell, C.: Invariant descriptors for 3d object recognition and pose. PAMI **13**(10) (1991) 971–991
14. Junejo, I., Dexter, E., Laptev, I., Perez, P.: Cross-view action recognition from temporal self-similarities. Technical report, Irisa, Rennes (2008) Publication interne N 1895, ISSN 1166-8687.
15. Farhadi, A., Kamali, M.: Learning to recognize activities from the wrong view point. In: ECCV. (2008)
16. Farhadi, A., Forsyth, D., White, R.: Transfer learning in sign language. In: CVPR. (2007)
17. Sadeghi, F., Zitnick, C.L., Farhadi, A.: Visalogy: Answering visual analogy questions. In Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., eds.: Advances in Neural Information Processing Systems 28. Curran Associates, Inc. (2015) 1882–1890
18. Reed, S.E., Zhang, Y., Zhang, Y., Lee, H.: Deep visual analogy-making. In Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., eds.: Advances in Neural Information Processing Systems 28. Curran Associates, Inc. (2015) 1252–1260
19. Ghifary, M., Bastiaan Kleijn, W., Zhang, M., Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders. In: The IEEE International Conference on Computer Vision (ICCV). (December 2015)
20. Tulsiani, S., Carreira, J., Malik, J.: Pose induction for novel object categories. In: The IEEE International Conference on Computer Vision (ICCV). (December 2015)
21. Friedman, J.H.: Stochastic gradient boosting. Computational Statistics and Data Analysis **38** (1999) 367–378

22. Ma, S., Huang, J.: Regularized roc method for disease classification and biomarker selection with microarray data. Bioinformatics **21**(24) (2005) 4356–4362
23. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: IEEE International Conference on on Robotics and Automation. (2011)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)