Totally Looks Like - How Humans Compare, Compared to Machines

Amir Rosenfeld , Markus D. Solbach , and John K.Tsotsos

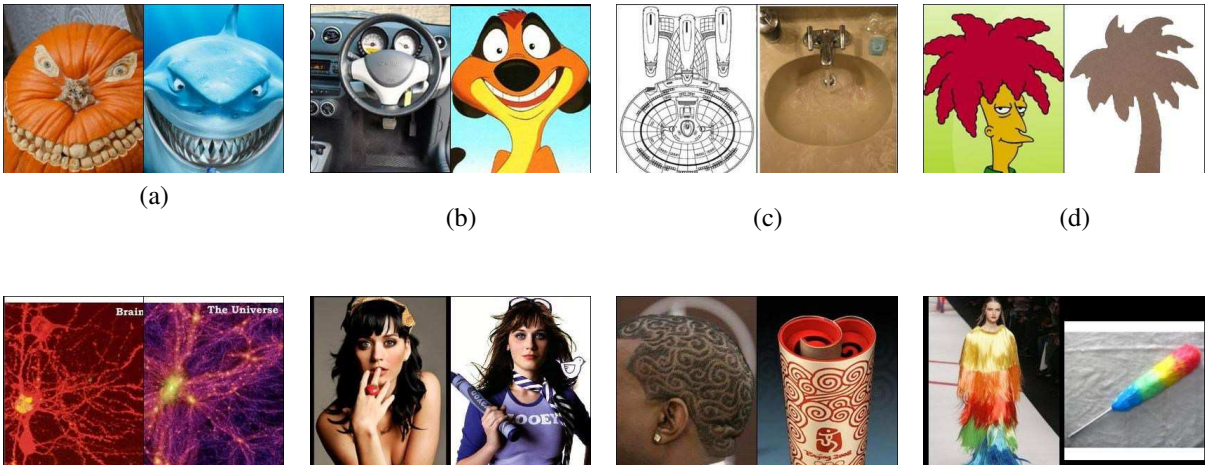
Department of Electrical Engineering and Computer Science
York University
Toronto, ON, Canada, M3J 1P3
{amir, solbach, tsotsos}@cse.yorku.ca

Abstract

Perceptual judgment of image similarity by humans relies on rich internal representations ranging from low-level features to high-level concepts, scene properties and even cultural associations. Existing methods and datasets attempting to explain perceived similarity use stimuli which arguably do not cover the full breadth of factors that affect human similarity judgments, even those geared toward this goal. We introduce a new dataset dubbed ***Totally-Looks-Like*** *(TLL) after a popular entertainment website, which contains images paired by humans as being visually similar. The dataset contains 6016 image-pairs from the wild, shedding light upon a rich and diverse set of criteria employed by human beings. We conduct experiments to try to reproduce the pairings via features extracted from state-of-the-art deep convolutional neural networks, as well as additional human experiments to verify the consistency of the collected data. Even though we create conditions to artificially make the matching task increasingly easier, we show that machine-extracted representations perform very poorly in terms of reproducing the matching selected by humans. The results suggest future directions for improvement of learned image representations. Data and code will be available at* https://sites.google.com/view/totally-looks-like-dataset.

1. Introduction

Human perception of images goes far beyond objects, shapes, textures and contours. Viewing a scene often elicits recollection of other scenes whose global properties or relations resemble the currently observed one. This relies on a rich representation in image space in the brain, entailing scene structure and semantics, as well as a mechanism to use the representation of an observed scene to recollect similar ones from the profusion of those stored in memory. In this work, we explore how representations based on deep neural networks fare on the challenge of similarity judgment between pairs of images from a new dataset, dubbed "**Totally-Looks-Like**" (TLL); See Figure 1. It is based on a website for entertainment purposes, which hosts pairs of images deemed by users to appear similar to each other, though they often share little common appearance, if judging by low-level visual features, Including objects, scenes, patterns, animals, and faces across various modalities (sketch, cartoon, natural images). Though not very large, the diversity and complexity of the images capture various aspects of human perception of image similarity, beyond current datasets. We evaluate the performance of several state-of-the-art models on this dataset, cast as a task of image retrieval. We compare this with human similarity judgments, forming not only a baseline for future evaluations, but also revealing specific weaknesses in the strongest of the current learned representations, pointing the way for future research. Human experiments validate the consistency of the data. Though in some experiments we allow favorable conditions for the machine-learned representations, they still often fall short of correctly predicting the human matches. Other lines of work also measure and analyze differences between human and machine perception, including [9, 4, 1, 8, 12]. These all show that human similarity measurements can be predicted quite well with modern, deep-learned representations. The proposed dataset shows where these methods fall short; our dataset is smaller in scale than most of them, but features images from the "wild", requiring similarities to be explained by features ranging from low-level to abstract scene properties. In this context, the proposed dataset does not contradict the systematic evaluations performed by prior art, but rather complements them and broadens the scope to see where modern image repre-



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1: The *Totally-Looks-Like* dataset: pairs of perceptually similar images selected by human users. The pairings shed light on the rich set of features humans use to judge similarity. Examples include (but are not limited to): attribution of facial features to objects and animals *(a,b)*, global shape similarity *(c,d)*, near-duplicates *(d)*, similar faces *(e)*, textural similarity *(f)*, color similarity *(g)*

sentations still lack.

2. Dataset

The data for the reported experiments is collected from a popular website called *TotallyLooksLike*[1]. The website describes itself simply as "Stuff That Looks Like Other Stuff". For the purpose of amusement, users can upload pairs of images which, in their judgment, resemble each other. Such images may have any content, such as company logos, household objects, art-drawing, faces of celebrities and others (cf. Figure 1). Little do most of the visitors of the website realize that it is in fact a hidden treasure: humans encounter an image in the wild and recall another image which not only do they deem similar, but so do hundreds of other site users (according to the votes). This provides a dataset of thousands of such image pairings, collected from the wild, that may aid to explore the cognitive drive behind judgment of image similarity. Beyond this, it contains samples of images that one recollects when encountering others, allowing exploration in the context of long-term visual memory and retrieval [5]. The collected dataset, *Totally-Looks-Like* (**TLL**), is a snapshot of 6016 image-pairs downloaded from the website in Jan. 2016, with permission from the web-site's administrators to make it publicly available for research purposes. We refer to the images in each pair as the "left" and the "right" images, or more concisely as $< L_i, R_i >, i \in 1 \ldots N$ where N is the size of the dataset.

[1] http://memebase.cheezburger.com/totallylookslike

3. Experiments

We wish to test to what degree similarity metrics based on generic machine-learned representations are able to reproduce the human-generated pairings in the TLL dataset. We frame this as an image retrieval task, where different deep-learned representations, both generic ([10, 3, 11, 6, 2] and learned for face recognition[2] are tested. This is done in two settings. In the first ("full retrieval"), all images are used as candidates and the distance (ℓ_2 or cosine distance, depending on the type of representation) is used to rank the images. In the second ("associative recall"), a small subset of images is sampled randomly (or according to some other criteria) and the computed features are used again to re-rank the images, to simulate a process of "associative recall", where a small, relevant subset of images is brought forth by some mechanism to be used as image candidates. This is to make the comparison more akin to human recollection, as humans likely do not perform exhaustive search but rather associate an observed image with a small subset of those stored in memory. It is an easier scenario compared to the full-retrieval one, simulating a state where a recall of a relevant dataset were available. The performance on the full retrieval task on a subset of the dataset which included roughly 1800 images using representations based on various deep-learned features in shown in Figure 2. The sub-

[2] https://github.com/ageitgey/face_recognition

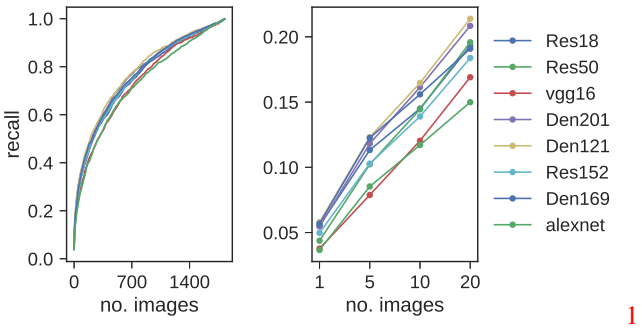
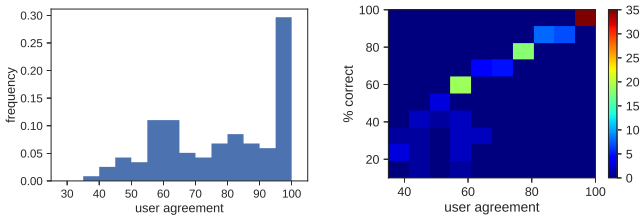


Figure 2: Retrieval performance by various learned representations in the TLL dataset. Left: all images. Right: showing recall only for the top 1 (first place), 5, 10, 20 images.

set was obtained by removing near-duplicate images (using both face and generic features) which rendered the dataset ambiguous. The performance on the simulation of "associative" memory is shown in Table 1 (a), where 10 runs were done for each condition (m, number of randomly picked images). The variance was negligible in these experiments, always $\leq \%1$.

3.1. Human Experiments

To compare humans against the machine-learned features, as well as verify the TLL dataset consistency, we performed human experiments on 120 images chosen using different criteria. The task was to select the best match given a query image and 5 putative candidates where the "correct" match is always present; we had 12 participants (ages 28-39) perform the experiment in-lab and 20 participants via Amazon Mechanical Turk (20). The 120 queries were split to 6 different conditions, involving whether the dataset includes only images of faces (TLL_d) or not (TLL_{obj}, mixed faces and other images). Based on this different distractor images are selecte for the humans. The distractor images are either chosen randomly or by sorting images based on their distance to the query images using generic features (densenet121), face features or generic features extracted from the face area. The results of the the machine and different human participants under different distractor types are indicated in Table 1 (b). Though not quite perfect, there is large consistency between the human workers on AMT and the users that uploaded the original TLL images. The performance of the lab-tested humans seems to be higher on average than the AMT workers, hinting that either the variability in human answers is rather large or that the AMT results contain some noise. Indeed, the number of votes given to each of the five options reveals trend to select the first option the most; the number of times each option was selected was 627, 522, 465, 395, 391; option 1 selected



(a) (b)

Figure 3: (a) Probability of agreement between human users on the AMT experiment. Humans tend to be highly consistent in their answers. (b) user agreement ratio vs. correct matching with TLL.

30% more times than the expected probability. Nevertheless, we see quite a high agreement rate throughout the table.

Human vs Machine Performance: the average human performance is generally lower when distractors are selected non-randomly, as expected. This is especially true for face images, where deep-learned features are used to select the distractor set; here AMT humans achieve around 60% agreement with the TLL dataset. This is not very surprising, as deep-learned face representations have already been reported to surpass human performance several years ago [7]. This may suggest that for faces, distractor images brought by the automatic retrieval seemed like better candidates to the humans than the original matches. The relatively high machine performance in the "random" cases is due to random distractors which were likely no closer in feature-space than the nearest neighbors of the query. We further show the consistency among human. We count for each query the frequency of each answer and test how many times humans agreed between themselves. In 87% of the cases, the majority of users (at least 11 out of 20) agreed on the answer. The most frequent event, (30%) was a total agreement of all users: 20 out of 20 repeated the same answer to the question. The Pearson correlation coefficient between user agreement and a correct match in TLL was 0.94. The plot of agreement percentage frequency is shown in Figure 3 (a). This large agreement is not in contradiction to the lower rates of success in reproducing the TLL results, because the TLL dataset was generated by a different process of unconstrained recollection, rather than forced choice as in our experiments. Figure 3 (b) shows the relation between user agreement ratios and the distribution of correctly answered images.

4. Conclusion

Deep-learned representations were shown to fall quite short of predicting the TLL image pairings, showing that human similarity perception is still not fully predicted by

m	% correct
1	100.00
2	73.35
3	61.54
4	54.30
5	50.49
10	37.99

(a)

	TLL_{obj}		TLL_d			
	random†	generic	random	face	face-generic	generic
human(lab)	83.3	70	82.5	63.3	64.5	83.3
human(AMT)	84	68.25	90.25	59	60.5	74.5
machine	20	20	25	0	0	5

(b)

Table 1: (a) Modeling Associative Recall: percentage of correct matches using conv-net derived features for the TLL dataset when a random sample of m images including the correct one is used. For 10 images, the performance is less that 50%. (b) man-versus-machine image matching accuracy for the perceptual similarity task. †The relatively high accuracy for "random" is because a small subset is selected which contains the correct answer, highly increasing the chance for correct guessing.

modern models. Despite some noise in the AMT data the statistics still clearly show humans to be quite consistent in choosing image pairs, even when faced confusing distractors. Emulating easier scenarios for machines (Table 1 (a)) yielded slightly improved results, but still far from reproducing the consistency observed among humans. One could argue that fine-tuning the machine learned representation with a subset of images in this dataset will reduce the observed gap. However, we believe that generic enough visual features should be able to reproduce the same similarity measurements without being explicitly trained to do so, just as humans do. Moreover, the set of various features employed by humans is likely rather large; previous attempts to reproduce human similarity measurements resulted in datasets much larger than the proposed one, though they were narrower in scope (cf [9]). This raises the question of the size of dataset required to close this gap in performance. We hypothesize that the set of features humans use for image comparison is not fixed, but conditional on the content of each pair of images. An ongoing work is to develop methods to apply such conditional computations to improve image based reasoning and representations. **Acknowledgement** This research was supported by several sources, via grants to the senior author, for which the authors are grateful: Air Force Office of Scientific Research USA (FA9550-18-1-0054), the Canada Research Chairs Program (950-231659), and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05352).

References

[1] Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Modeling human categorization of natural images using deep feature representations. *arXiv preprint arXiv:1711.04855*, 2017. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[3] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016. 3

[4] Kamila M. Jozwik, Nikolaus Kriegeskorte, Katherine R. Storrs, and Marieke Mur. Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. *Frontiers in Psychology*, 8:1726, 2017. 1

[5] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and Predicting Image Memorability at a Large Scale. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[7] Chaochao Lu and Xiaoou Tang. Surpassing Human-Level Face Verification Performance on LFW with GaussianFace. In *AAAI*, pages 3811–3819, 2015. 3.1

[8] Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*, 2016. 1

[9] RT Pramod and SP Arun. Do computational models differ systematically from human object perception? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1601–1609, 2016. 1, 4

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv preprint arXiv:1801.03924*, 2018. 1