Totally Looks Like - How Humans Compare, Compared to Machines[*]

Amir Rosenfeld, Markus D. Solbach, and John K. Tsotsos
{amir, solbach, tsotsos}@cse.yorku.ca

York University
Toronto, ON, Canada, M3J 1P3

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. However, 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. 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. We discuss and analyze these results, suggesting future directions for improvement of learned image representations.

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 of 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. Though not

Fig. 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 *(e)*, similar faces *(f)*, textural similarity *(g)*, color similarity *(h)*

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

(e)   (f)   (g)   (h)
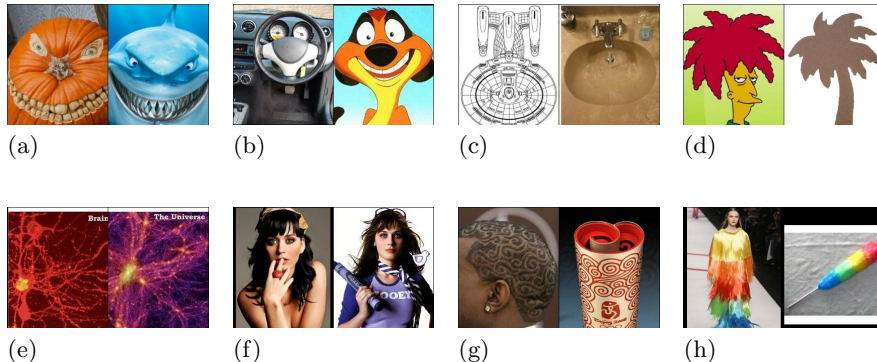
fully understood, the capacity of the human brain to memorize images is surprisingly large [3,12]. The recent explosion in the performance and applicability of deep-learning models in all fields of computer vision [19,25,14] (and others), including image retrieval and comparison [26], can tempt one to conclude that the representational power of such methods approaches that of humans, or perhaps even exceeds them. We aim to explore this by testing how 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. These include pairs of images out of (but not limited to) objects, scenes, patterns, animals, and faces across various modalities (sketch, cartoon, natural images). The website also includes user ratings, showing the level of agreement with the proposed resemblances. Though it is not very large, the diversity and complexity of the images in the dataset implicitly captures many aspects of human perception of image similarity, beyond current datasets which are larger but at the same time narrower in scope. 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 that point the way for future research and improvements. We conduct human experiments to validate the consistency of the collected data. Even though in some experiments we allow very favorable conditions for the machine-learned representations, they still often fall short of correctly predicting the human matches.

The next section overviews related work. This is followed by a description of our method, experiments and analysis. We close the paper with discussion about the large gaps between what is expected of state-of-the art learned representations and suggestions for future work. The dataset is available at the following address: `https://sites.google.com/view/totally-looks-like-dataset`

2 Related Work

This paper belongs to a line of work that compares machine and human vision (in the context of perception) or attempts to perform some vision related task that is associated with high-level image attributes. As ourselves, others also tapped the resources of social media/online entertainment websites to advance research in high-level image understanding. For example, Deza and Parikh [6] collected datasets from the web in order to predict the virality of images, reporting super-human capabilities when five high-level features were used to train an SVM classifier to predict virality.

Several lines of work measure and analyze differences between human and machine perception. The work of [17] collected 26k perceived dissimilarity measurements from 2,801 visual objects across 269 human subjects. They found several discrepancies between computational models and human similarity measurements. The work of [10] suggests that much of human-perceived similarity can readily be accounted for by representations emerging in deep-learned models. Others modify learned representations to better match this similarity, reporting a high-level of success in some cases [16], and near-perfect in others [2]. The work of [2] is done in a context which reduces similarity to categorization. Very recently, Zhang Et al. [24] have shown that estimation of human perceptual similarity is dramatically better using deep-learned features, whether they are learned in a supervised or unsupervised manner, than more traditional methods. Their evaluation involved comparing images to their distorted versions. The distortions tested were quite complex and diverse. Akin to ours, there are works who question the behavioral level of humans vs. machines. For instance, Das et. al [5] compare the attended image regions in Visual Question Answering (VQA, [1]) to that of humans and report a rather low correlation. Other works tackle high level tasks such as understanding image aesthetics [22] or even humor [4]. The authors of [7] compare the robustness of humans vs. machines to image degradations, showing that DNN's that are not trained on noisy data are more error-prone than humans, as well as having a very different distribution of non-class predictions when confronted with noisy images. Matching images and recalling them are two very related subjects, as it seems unlikely for a human (or any other system storing a non-trivial amount of images) to perform exhaustive search over the entire collection of images stored in memory. Studies of image memorability [11] have successfully produced computational models to predict which images are more memorable than others.

The works of [17,16,10,24] show systematic results on large amounts of data. However, most of the images within them either involve objects with a blank

background [17,10] or of a narrow type (e.g., animals [16]). Our dataset is smaller in scale than most of them, but it features images from the "wild", requiring similarities to be explained by features ranging from low-level to abstract scene properties. In [24], a diverse set of distortions is applied to images, however, the source image always remains the same, whereas the proposed dataset shows pairs of images of different scenes and objects, still deemed similar by human observers. 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 representations still fall short.

3   Method

The main source of data for the reported experiments is 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 be have any content, such as company logos, household objects, art-drawing, faces of celebrities and others. Figure 1 shows a few examples of such image pairings. Each submission is shown on the website, and viewers can express their agreement (or disagreement) about the pairing by choosing to up-vote or down-vote. The total number of up-votes and down-votes for each pair of images is displayed.

Little do most of the casual visitors of this humorous 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, by definition 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.

While other works have explored image memorability [11], in this work we focus on the aspects of similarity judgment. We next describe the dataset we created from this website.

3.1   Dataset

We introduce the *Totally-Looks-Like* (**TLL**) dataset. The dataset contains a snapshot of 6016 image-pairs along with their votes downloaded from the website in Jan. 2018 (a few images are added each day). The data has been downloaded with permission from the web-site's administrators to make it publicly available for research purposes. For each image pair, we simply refer to the two images as the "left image" and the "right image", or more concisely as $< L_i, R_i >, i \in 1 \ldots N$ where N is the total number of images in the dataset. We plan to make the data available on the project website, along with pre-computed features which will be listed below.

---

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

3.2   Image Retrieval

The TLL dataset is the basis for our experiments. We wish to test to what degree similarity metrics based on generic machine-learned representations are able to reproduce the human-generated pairings.

We formulate this as a task of image retrieval: Let $\mathcal{L} = (L_i)_i$ be the set of all left images and similarly let $\mathcal{R}$ be the set of all right images. For a given image L_i we measure the distance $\phi(L_i, R_j)$ between L_i and each $R_j \in \mathcal{R}$. This induces a ranking $r_1, \ldots r_n$ over $\mathcal{R}$ by sorting according to the distance $\phi(\cdot, \cdot)$. A perfect ranking returns $r_1 = i$. Calculating distances using ϕ over all pairs of the dataset allows us to measure its overall performance as a distance metric for retrieval. For imperfect rankings, we can measure the recall up to some ranking k, which is the average number of times the correct match was in the top-k ranked images. In practice, we measure distances between feature representations extracted via state-of-the-art DCNN's, either specialized for generic image categorization or face identification, as detailed in the experiments section.

Direct Comparison vs. Recollection: We note that framing the task as image retrieval may be unfair to both sides: when humans encounter an image and recollect a perceptually similar one to post on the website, they are not faced with a forced choice task of selecting the best match out of a predetermined set. Instead, the image triggers a recollection of another image in their memory, which leads to uploading the image pair. On one hand, this means that the set of images from which a human selects a match is dramatically larger than the limited-size dataset we propose, so the human can potentially find a better match. On the other hand, the human does not get to scrutinize each image in memory, as the process of recollection likely happens in an associative manner, rather than by performing an exhaustive search on all images in memory. In this regard, the machine is more free to spend as many computational resources as needed to determine the similarity between a putative match. Another advantage for the machine is that the "correct" match already exists in the predetermined dataset; possibly finding it will be easier than in an open-ended manner as a human does. Nevertheless, we view the task of retrieval from this closed set as a first approximation. In addition, we suggest below some ways to make the comparison more fair.

4   Experiments

We now describe in detail our experiments, starting from data collection and preprocessing, through various attempts to reproduce the human data and accompanying analysis.

Data Preprocessing All images if the TLL (Totally-Looks-Like) dataset were automatically downloaded along with their up-votes and down-votes from the website. Each image pair $< L_i, R_i >$appears on the website as a single image showing L_i and R_i horizontally concatenated, of constant width of 401 pixels and height of 271 pixels. We discard for each image the last column and split it

equally to left and right images. In addition, the bottom 26 pixels of each image contains for each side a description of the content. While none of the methods we apply explicitly use any kind of text detection/recognition, we discard these rows as well to avoid the possibility of "cheating" in the matching process.

4.1   Feature extraction

We extract two kinds of features from each image: generic and facial.

Generic Features: we extract "generic" image features by recording the output of the penultimate layer of various state-of-the-art network architectures for image categorization, trained on the ImageNet benchmark [18], which contains more than a million training images spread over a thousand object categories. Training on such a rich supervised task been shown many times to produce features which are transferable across many tasks involving natural images [20]. Specifically, we use various forms of Residual Networks [8], Dense Residual Networks,[9], AlexNet [13] and VGG-16 from [21], giving rise to feature-vector dimensionalities ranging from a few hundred to a few thousands, dependent on the network architecture. We extract the activations of the penultimate layer of each of these networks for each of the images and store them for distance computations.

Facial Features: many of the images contain faces, or objects that resemble faces. Faces play an important role in human perception and give rise to many of the perceived similarities. We run a face detector on all images, recording the location of the face. For each detected face in each image, we extract features using a deep neural network which was specifically designed for face recognition. The detector and features both use an off-the-shelf implementation [2]. The dimensionality the extracted face descriptor is 128. Figure 5 (c) shows the distribution of the number of detected faces in images, as well as the agreement between the number of detected faces in human-matched pairs. The majority of images have a face detected in them, which very few containing more than one face. When a face is detected in a left image of a given pair, it is likely that a face will be detected in the right one as well.

Generic-Facial Features : very often in the TLL dataset, we can find objects that resemble faces and play an important role in these images, being the main object which led to the selection of an image pair. To allow comparing such objects to one another, we extract generic image features from them, as described above, to complement the description by specifically tailored facial features. We do this under the likely assumption that while a facial feature extractor might not produce reliable features for comparison from a face-like object (because the network was not trained on such images), a generic feature extractor might.

We denote by G_i, F_i, and GF_i the set of generic features, facial features and generic-facial features extracted from each image. Note that for some images faces are not at all detected, and so F_i and GF_i are empty sets. For others,

---

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

possibly more than one face is detected, in which F_i and GF_i can be sets of features.

We next describe how we take all of these features into account.

4.2   Matching Images

We define the distance function between a pair of images L_i, R_j by their extracted features as described above. We either use the ℓ_2 (Euclidean) distance between a pair of features, i.e., $\phi_l^f(A,B) = \|A - B\|_2$ or the cosine distance, i.e, $\phi_c^f(A,B) = 1 - \frac{A \cdot B}{\|A\|\|B\|}$. Where A, B are the corresponding features for images L_i, R_j. The subscripts l, c specify ℓ_2 norm or cosine distance. The superscript f specifies the kind of representation used, i.e, $f \in \{G, F, GF\}$. For facial features (F) we use only the euclidean distance, as is designated by the applied facial recognition method. Each distance function ϕ_l^f generates a distance matrix $\Phi_l^f \in \mathcal{R}^N$ with the i, j location representing the distance between L_i, R_j using this function. For image pairs with more than one face in either image we assign the corresponding i, j location the minimal distance between all pairs of features extracted from the corresponding faces. For image pairs where at least one image has no detected face we assign the corresponding distance to $+\infty$.

Armed with Φ_l^f, we may now test how the distance-induced ranking aligns with the human-selected matches.

Evaluating Generic Features: as a first step, we evaluate which metric (Euclidean vs. cosine) better matches the pairings in TLL. We noted that the recall for a given number of candidates using the cosine distance is always higher compared to that of the Euclidean distance. This can be seen in Figure 2 (a). We calculated recall for each of the nets as a function of the number of retrieved candidates. The figure shows the difference for each k between the recall for the cosine vs. Euclidean distances. The cosine distance has a clear advantage here, hence we choose to use it for all subsequent experiments (except for the case of facial features).

Near duplicates: visualizing some of the returned nearest neighbors revealed that there are duplicate (or near duplicate images) within the $\mathcal{L}$ and $\mathcal{R}$ image sets. As this could cause an ambiguity and hinder retrieval scores, we removed all pairs where either the left or the right image was part of the duplicate. We did this for both generic features and face-based features. For generic ones, this corresponds to a cosine distance of ≥ 0.15 (using Densenet121); virtually all images below a distance of 0.1 were near-duplicate, so we set the threshold conservatively to avoid accidental duplicates. For faces we set the threshold to 0.5. We also removed duplicates across pairs, meaning that if L_i and R_j were found to be near-duplicates then we removed them, as an identical copy R_j of L_i may be a better match for it than R_i. Removing all such duplicates leaves us with a subset we name TLL_d, containing 1828 valid image pairs. The results of Table 1 and 2(a) are calculated based on this dataset. This does not, however, reduce the importance of the full dataset of 6016 images as it still contains many

Fig. 2: (a) Difference between recall per number of images retrieved for cosine and ℓ_2-distance based retrieval. Recall is always improved if we use the cosine distance over the ℓ_2 distance between representations. (b) 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.
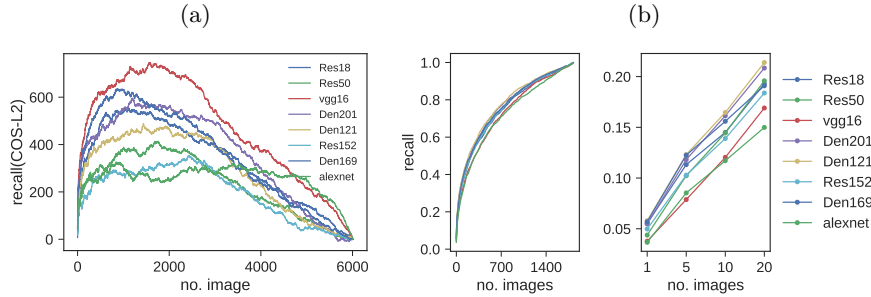
interesting and useful image pairs to learn from. The reduction of the dataset size is only done for evaluation purposes.

Faces: many images in the dataset contain faces, as indicated by Figure 5 (c). In fact, the figure represents an underestimation of the number of faces as some faces we not detected. Such images seem qualitatively different from the ones containing faces, in that the similarities are more about global shape, texture, or face-like properties, though there are no actual faces in them in the strict sense. Hence, we create another partition of the data without any detected faces, and without the duplicate images according to the generic feature criteria. This subset, TLL_{obj}, contains 1622 images. Both TLL_d and TLL_{obj} are used in Section 4.3 where we report additional results of human experiments.

	R@1	R@5	R@10	R@2	R@50	R@100
AlexNet	3.67	9.19	12.09	15.37	22.59	30.63
vgg16	3.77	8.97	12.58	16.90	24.02	32.39
Res50	4.38	11.43	15.04	19.91	28.77	36.71
Res152	4.98	11.16	14.61	18.82	26.20	35.61
Den201	5.47	12.91	16.63	21.44	30.47	38.18
Res18	5.53	12.14	15.10	19.47	28.06	35.61
Den169	5.69	13.07	16.19	19.31	28.67	37.53
Den121	5.80	13.84	16.90	21.94	29.92	38.89

Table 1: Retrieval performance (percentage retrieved after varying number of candidates) by various learned representations in the TLL dataset.

Next, we evaluate the retrieval performance as a function of the number of returned image candidates. This can be seen graphically in Figure 2 (b). The left sub-figure shows the recall for the entire dataset and the right sub-figure shows it for the first, 5th, 10th and 20th returned candidates. Table 1 shows these values numerically. For face features the retrieval accuracy using one retrieve item was slightly better than the generic features, reaching 6.1%. Using generic features extracted on faces performed quite poorly, at 2.6%. Evidently, none of the networks we tested performed well on this benchmark. Such a direct comparison is problematic for several reasons. Next, we attempt to ease the retrieval task for the machine-based features.

Simulating Associative Recall : As mentioned in Sec. 3.2, directly comparing to all images in the dataset is perhaps unfair to the machine-learning test. Arguably, a human recalling an image first narrows down the search given the query image, so only images with relevant features are retrieved from memory. Though we do not speculate about how this may be done, we can test how retrieval improves if such a process were available. To do so, we sample for each left image L_i a random set $R(L_i)$ of size m which includes the correct right image R_i and an additional $m-1$ images. This simulates a state where viewing the image L_i elicited a recollection of m candidates (including the correct one) from which the final selection can be made. We do this for varying sizes of a recollection set $m \in \{1-5, 10, 20, 50, 100\}$, with 10 repetitions each. Table 2 (a) summarizes the mean performance obtained here. Although these are almost "perfect" conditions, the retrieval accuracy falls to less than 50% if we use as little as ten examples as the test set. The variance (not shown) was close to 0 in all conditions.

Comparing Distances to Votes: we test whether there is any consistency between the feature-based distances and the number of votes assigned by human users. Assuming that a similar number of users viewed each uploaded image pair, a higher number of votes suggests higher agreement that the pairing is indeed a valid one. Possibly, this could also suggest that the images should be easier to match by automatically extracted features. We calculate the correlation between number of up-votes and down-votes vs the cosine-distance resulting from the Densenet121 network. Unfortunately, there seems very little correlation, with a Pearson coefficient of 0.023 / -0.068 for up/down-votes respectively. Hence the following experiments do not use the voting information.

4.3   Human Experiments

We conducted experiments both in-lab and using Amazon Mechanical Turk (AMT). We chose 120 random pairs of images from the dataset, as follows: 40 pairs were selected TLL_{obj} and 80 from TLL_d. From each pair, we displayed the left image to the user, along with 4 additional selected images and the correct right image. The images were shuffled in random order. Human subjects were requested to select the most similar image to the query (left) image. We allocated 20 images to each sub-experiment. The names of the experiments are **random**, **generic**, **face** and **face-generic**, indicating the type of features used

Fig. 3: Automatic retrieval errors: using distances between state-of-the-art deep-learned representations often does not do well in reproducing human similarity judgments. Each row shows a query image on the left, five retrieved images and the ground-truth on the right. Perceptual similarity can be attributed to similarity between cartoonish and real faces (first three rows), flexible transfer of facial expression (4th row), visually similar sub-regions (last two rows, hair of person on row 5 resembles spider legs, hair of person on last row resembles waves). Though the images and the retrieved ones may be much more similar to each other in a strict sense, humans still consistently agree on the matched ones (first, last columns).

to select the subset, if any. For **random** we simply chose a subset of 5 images randomly, similarly to what is described in 4.2. For each of the others, we ordered the images from the corresponding subset using each feature type and retained the top-5. If the top-5 images retrieved did not contain the correct answer, we randomly replaced one of them with it. A correct answer in this sense is selecting the correct right image, for the human, and ranking it highest for the machine. In each experiment, the four images except the correct match are regarded as *distractors*. Distractors generated using feature similarity (as opposed to random selection) pose a greater challenge for human participants, as they tend to resemble, in some sense, the "correct" answer. Table 2 (b) summarizes the overall accuracy rates. In lab settings (12 participants, ages 28-39) answered all 120 questions each (labeled human1,human2 in the table). For AMT, we repeated each experiment 20 times, where an experiment is answering a single query, making an overall of 2400 experiments. A payment of 5 cents was rewarded for the

m	% correct
1	100.00
2	73.35
3	61.54
4	54.30
5	50.49
10	37.99
20	27.23
50	13.37

(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 2: (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.

completion of each experiment. Only "master" workers were used in the experiment, for increased reliability. We next highlight several immediate conclusions from this data.

Data Verification: the first utility of the collected human data is to validate the consistency of that collected from the website. 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, when we count the number of votes given to each of the five options, we note a trend to select the first option the most, persisting through options 2-4. The number of times each option was selected was 627, 522, 465, 395, 391; option 1 selected 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 [15]. 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 very low consistency of the machine retrieval with humans is consistent with what is reported in table 1; the less than 6% performance rates translated to 0, in this specific sample of twenty examples for each test case. The relatively high per-

formance in the "random" cases is due to selection of random distractors which were likely no closer in feature-space than the nearest neighbors of the query, hence resulting in seemingly high performance. We further show the consistency among human users by counting the number of agreements on answers. 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. In fact, the most frequent event, occurring 30% of the time, was a total agreement - 20 out of 20 identical answers. Moreover, the Pearson correlation coefficient between user agreement and a correct matching to TLL was 0.94. The plot of agreement frequencies is shown in Figure 5 (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 5 (b) shows the relation between user agreement ratios and the distribution of correctly answered images.

Finally, Figure 4 shows four queries from the dataset, in the form of one query image (left column) and five candidates (remainder columns). Two of the rows shows cases where there was a perfect human agreement and two show cases where the answers were almost uniformly spread over the candidates. It is not difficult to guess which rows represent each case.

Fig. 4: Sample queries with varying user agreement. Each row shows on the left column a query image and 5 images from which to select a match. Some queries are very much agreed upon and on some the answers are evenly distributed. We show two rows of the first case, and two of the second. We encourage the reader to guess which images were of each kind.
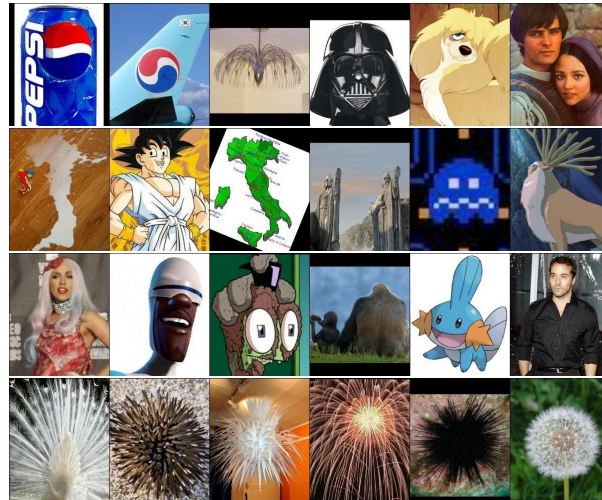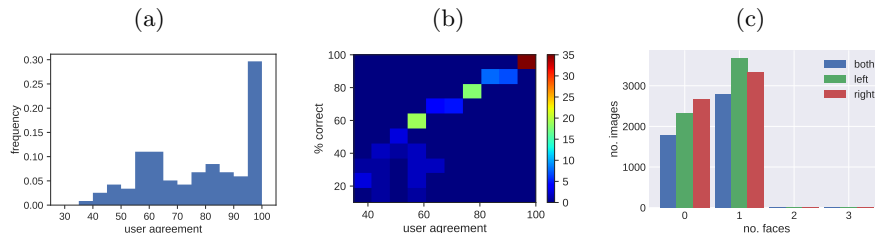
Fig. 5: (a) Probability of agreement between human users on the AMT experiment. Humans tend to be highly consistent in their answers. (b) user agreement ratio vs. correct matching with TLL. (c) Distribution of number of detected faces and agreement on detected faces between left-right image pairs.


5   Discussion

We have looked into a high-level task of human vision: perceptual judgment of image similarity. The new TLL dataset offers a glimpse into images which are matched by human beings in the "wild", in a less controlled fashion, but arguably one that sheds a different light on various factors compared previous work in this area. Most works in image retrieval deal with near-duplicate images, or images which mostly depict the same type of concept. We explored the ability of existing state-of-the-art deep-learned features to reproduce the matchings in the dataset. Though one would predict this to produce a reasonable baseline, neither features resulting from object classification networks and ones tailored for face verification seem to be able to remotely reproduce the matchings between the image pairs. We verified this using additional human experiments, both in-lab and using Amazon Mechanical Turk. Tough the collected data from AMT was not cleaned and clearly showed signs of existence of biases, the statistics still clearly show that humans are quite consistent in choosing image pairs, even when faced with a fair amount of distractors. Emulating easier scenarios for machines (for example, Table 2 (a)) yielded improved results, but ones which are still very 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 sufficiently generic 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 in terms of image variability (for example [17]). This raises the question, how many images will an automatic method require to reproduce this rich set of similarities demonstrated by humans?

Fig. 6: Additional examples. Perceived image similarities can be abstract/symbolic: cats $\leftrightarrow$ guards, doorway $\leftrightarrow$ mountain passageway *(a)*, low-level (colors, *(d,e,f)*, 2D shape *(b,c,e,g)*, 3D-shape *(e)*, related to well-known iconic images from pop-culture *(b,e,f,h)* art *(c)* or pose-transfer across very different objects/domains *(b,c,d)*

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

(e)   (f)   (g)   (h)
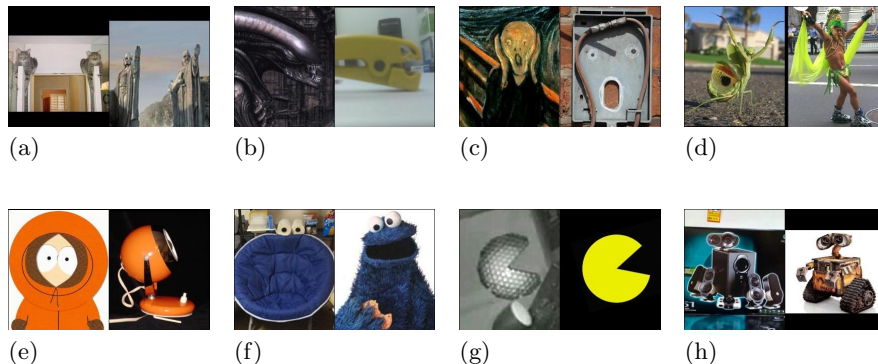
We do not expect strong retrieval systems to reproduce the matchings in TLL. On the contrary, a cartoon figure should not be automatically associated with the face of Nicolas Cage 3 (2nd row), this would likely constitute a retrieval error in normal conditions and lead to additional unexpected ones. However, we do expect a high-level representation to report that of all the images in that row, the most similar one is indeed that of the said actor. Humans can easily point to the facial features in which the cartoon and the natural face image bear resemblance. In fact, we believe that for similarity judgments to be consistent with those of humans (note there is no "correct" or "incorrect"), they should be multi-modal and conditioned on *both* images. Relevant factors include (1) facial features (2) facial expressions (3rd row in Figure 3), requiring a robust comparison between facial expressions in different modalities (3) texture or structure of part of the image (last row, person's hair). The factors are not fixed or weighted equally in each case. Additional factors involve comparison between different objects or familiarity with iconic images or characters as depicted in Figure 6.

As the importance of factors changes as a function of the image-pair, we suggest that the comparison will be akin to visual-question-answering (VQA) , in the form "*why should image A be regarded as similar / dissimilar to image B?*". Just as VQA models on single images benefit from attention models [23], we suggest that asking a question that requires extracting relevant information from two different images will give rise to attention being applied to both. Information extracted from one image (such as the presence of a face, waves, an unusual facial expression, or spider-legs in Figure 3) is necessary to produce a basis for comparison and feature extraction from the other. We leave further development of this direction to future work.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
2. Battleday, R.M., Peterson, J.C., Griffiths, T.L.: Modeling human categorization of natural images using deep feature representations. arXiv preprint arXiv:1711.04855 (2017)
3. Brady, T.F., Konkle, T., Alvarez, G.A., Oliva, A.: Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences **105**(38), 14325–14329 (2008)
4. Chandrasekaran, A., Vijayakumar, A.K., Antol, S., Bansal, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: We are humor beings: Understanding and predicting visual humor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4603–4612 (2016)
5. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding **163**, 90–100 (2017)
6. Deza, A., Parikh, D.: Understanding image virality. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1818–1826 (2015)
7. Geirhos, R., Janssen, D.H., Schütt, H.H., Rauber, J., Bethge, M., Wichmann, F.A.: Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)
10. Jozwik, K.M., Kriegeskorte, N., Storrs, K.R., Mur, M.: Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. Frontiers in Psychology **8**, 1726 (2017). https://doi.org/10.3389/fpsyg.2017.01726, `https://www.frontiersin.org/article/10.3389/fpsyg.2017.01726`
11. Khosla, A., Raju, A.S., Torralba, A., Oliva, A.: Understanding and Predicting Image Memorability at a Large Scale. In: International Conference on Computer Vision (ICCV) (2015)
12. Konkle, T., Brady, T.F., Alvarez, G.A., Oliva, A.: Scene memory is more detailed than you think: The role of categories in visual long-term memory. Psychological Science **21**(11), 1551–1556 (2010)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
14. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (2017)
15. Lu, C., Tang, X.: Surpassing Human-Level Face Verification Performance on LFW with GaussianFace. In: AAAI. pp. 3811–3819 (2015)
16. Peterson, J.C., Abbott, J.T., Griffiths, T.L.: Adapting deep network features to capture psychological representations. arXiv preprint arXiv:1608.02164 (2016)

17. Pramod, R., Arun, S.: Do computational models differ systematically from human object perception? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1601–1609 (2016)
18. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
19. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)
20. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 806–813 (2014)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Workman, S., Souvenir, R., Jacobs, N.: Quantifying and Predicting Image Scenicness. arXiv preprint arXiv:1612.03142 (2016)
23. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv preprint arXiv:1801.03924 (2018)
25. Zhou, P., Feng, J.: The Landscape of Deep Learning Algorithms. arXiv preprint arXiv:1705.07038 (2017)
26. Zhou, W., Li, H., Tian, Q.: Recent Advance in Content-based Image Retrieval: A Literature Survey. arXiv preprint arXiv:1706.06064 (2017)