"Are Machines Better Than Humans in Image Tagging?" - A User Study Adds to the Puzzle

Ralph Ewerth[1,2(✉)], Matthias Springstein[1], Lo An Phan-Vogtmann[3], and Juliane Schütze[3]

[1] German National Library of Science and Technology (TIB), Hannover, Germany
{ralph.ewerth,matthias.springstein}@tib.eu
[2] L3S Research Center, Hannover, Germany
[3] Jena University of Applied Sciences, Jena, Germany
loan.phan-vogtmann@stud.eah-jena.de, juliane.schuetze@eah-jena.de

Abstract. "Do machines perform better than humans in visual recognition tasks?" Not so long ago, this question would have been considered even somewhat provoking and the answer would have been clear: "No". In this paper, we present a comparison of human and machine performance with respect to annotation for multimedia retrieval tasks. Going beyond recent crowdsourcing studies in this respect, we also report results of two extensive user studies. In total, 23 participants were asked to annotate more than 1000 images of a benchmark dataset, which is the most comprehensive study in the field so far. Krippendorff's α is used to measure inter-coder agreement among several coders and the results are compared with the best machine results. The study is preceded by a summary of studies which compared human and machine performance in different visual and auditory recognition tasks. We discuss the results and derive a methodology in order to compare machine performance in multimedia annotation tasks at human level. This allows us to formally answer the question whether a recognition problem can be considered as solved. Finally, we are going to answer the initial question.

1 Introduction

In the field of multimedia analysis and retrieval, human performance in recognition tasks was reported from time to time [2,9,12,13,15,16,20–23], but has not been evaluated in a consistent manner. As a consequence, the quality of human performance is not exactly known and estimates exist only for few recognition tasks. The design of the related human experiments also varies noticeably in many respects. For example, crowdsourcing is often utilized to employ annotators [9,12–16,23], which is coming along with some methodological issues. The number of human participants varies from 1 to 40 in the studies considered in this paper. The same is true for the experimental instructions and their expertise, in particular for crowdworkers. This, for example, makes it nearly unfeasible to evaluate and compare machine performance at human level across different tasks. In fact, we know little *in general* about human performance in multimedia

J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193, pp. 186–198, 2017.
DOI: 10.1007/978-3-319-56608-5_15

content analysis tasks. As a consequence, the question when such a task can be considered as solved cannot be answered easily. The related question is addressed by this paper: How can we systematically set machine performance in relation to human performance? If human ground truth data are the (only) baseline, machine performance can basically never be better than (human) ground truth data. But considering the impressive recent advances in deep learning for pattern recognition tasks, it is desirable to set machine performance in relation to human-level performance in a systematic manner.

Another issue is related to ground truth data for retrieval tasks: The relevance of multimedia documents at retrieval time for a certain user is not known in advance and it depends on the user's current search task and context. The issue of evaluating multimedia analytics systems has been also stressed recently by Zahálka et al. [25]. For example, a detective is interested in every occurrence of a suspicious object (e.g., a car) in any size. On the other hand, a TV journalist, who searches for material for re-use in order to illustrate the topic mobility, might be interested only in retrieval results showing a car in an "iconic" view, i.e., placed clearly in the foreground.

In this paper, we review a number of papers reporting human performance in visual and auditory recognition tasks. This aims at putting together some parts of the puzzle: How well do machines perform in such tasks compared to humans? To answer this question, the results are set in relation to the current state of the art of automatic pattern recognition systems. Furthermore, we present a comprehensive user study that closes the gap of comparing in detail human and machine performance in annotation tasks for realistic images, as they are used in the PASCAL VOC (Visual Object Classes) challenge [4,5], for example. More than 1000 images have been annotated by 23 participants in a non-crowdsourcing setting. The number of images also allows us to draw conclusions about rarely occuring concept categories such as "cow" or "potted plant". It is suggested to evaluate the reliability of users' annotations by Krippendorff's α [10,11], which measures the agreement among several coders. The results of the presented study are discussed and conclusions are drawn for the evaluation of computer vision and multimedia retrieval systems: A methodology is introduced that enables researchers to formally compare machine performance at human level in visual and auditory recognition tasks. To summarize, the contributions of this paper are as follows:

- Surveying and comparing human and machine performance in a number of visual and auditory pattern recognition tasks,
- presenting a comprehensive user study regarding image annotation yielding insights into the relation of human and machine performance,
- introducing the concept of inter-coder reliability in the field of multimedia retrieval evaluation for comparing human and machine performance,
- proposing an evaluation methodology that allows us to evaluate machine performance at human level in a systematic manner, and
- suggesting two indices for measuring human-level performance of systems.

The remainder of the paper is structured as follows. Section 2 surveys studies that compared human and machine performance for different visual and auditory recognition tasks. Section 3 deals with a comprehensive user study regarding image annotation and related results are presented. A methodology to evaluate machine performance at human level in a systematic way is suggested in Sect. 4. Finally, some conclusions are drawn in Sect. 5.

2 Human and Machine Performance in Visual and Auditory Recognition Tasks

In this section, we briefly survey related work which compared human and machine performance for some multimedia analysis tasks. Yet, human performance has been considered only in a small number of studies.

Some studies evaluated human performance in the task of visual concept classification. Kumar et al. [12] as well as Lin et al. [14] measured the human inter-coder agreement and compared experts against crowdsourcing annotators. Other papers reported the performance of humans and machine systems on some benchmark datasets. Jiang et al. [9] presented a dataset for consumer video understanding. For this dataset, the human annotations were significantly better than the machine results. Parikh and Zitnick [16] investigated the role of data, features, and algorithms for visual recognition tasks. An accuracy of nearly 100% is reported for humans on two PASCAL VOC datasets, whereas machine performance was around 50% on both datasets in 2008. Xiao et al. [23] presented the dataset SUN (Scene Understanding) for scene recognition consisting of 899 categories and 130,519 images. Scene categories are grouped in an overcomplete three-level tree. Human performance reached 95% accuracy at the (easy) first level and 68.5% at the third level of the hierarchy, while the machine performance of 38% accuracy was significantly below human accuracy in this study. Russakovsky et al. [18] surveyed the advances in the field of the ImageNet challenge [3] from 2010 to 2014. The best result in 2014 was submitted by Szegedy et al. [19] (GoogLeNet) and achieved an error rate of 6.66%. Russakovsky et al. compared this submission with two human annotators and discovered that the neuronal network outperformed one of them. He et al. [8] claimed their system to be the first one that surpassed human-level performance (5.1%) on ImageNet data by achieving an error rate of 4.94%.

Phillips et al. [17] conducted one of the first comparisons of face identification capabilities of humans and machines. Interestingly, at that time the top three algorithms were already able to match or to do even better face identification compared with human performance on unfamiliar faces under illumination changes. Taigman et al. [20] presented a deep learning system for face verification that improves face alignment based on explicit 3D modelling of faces. A human-level performance of 97.35% accuracy was reported for the benchmark "Faces in the Wild" (humans: 97.53% accuracy).

Other interesting comparisons between humans and machine systems include camera motion estimation (Bailer et al. [2]), music retrieval (Turnbull et al. [20]),

and the geolocation estimation of images (Weyand et al. [22]). These studies also demonstrated that machines can reach or outperform human performance.

While the experiments of Parikh and Zitnick [16] with respect to re-engineering the recognition process did not provide evidence that humans are superior to machines, other reported results on PASCAL VOC and other data sets showed that humans perform significantly better on classifying natural scene categories. The reported results for visual concept classification [9,12,15,23] indicate that human performance is (far) better than the respective state of the art for automated visual concept classification at that time. Although He et al. claimed in 2015 that human-level performance has been surpassed by their approach [8], this claim remains questionable since only two human annotators were involved in the underlying study of Russakovsky et al. [18]. There are also some methodological issues in the reported experiments of the other studies, for example, experimental settings are not well defined, the employment of crowdsourcing is critical, or the number of images is too small which prevents drawing conclusions for rare classes. Hence, stronger empirical evidence still has to be provided for a meaningful comparison of human and machine performance. Therefore, in the remainder of this paper we address these issues by a comprehensive user study and derive a methodology to measure machine performance at human level.

3 User Study: Human Performance in Image Annotation

The analysis of previous work shows that the settings of the majority of studies do not allow us to compare human performance against machine learning approaches for image classification tasks. In this section, we present two user studies measuring human performance in annotating common image categories of daily life in a realistic photo collection (PASCAL VOC [4,5]). The design of this study is described in Sect. 3.1. The inter-coder agreement of the two experiments is evaluated using Krippendorff's α (Sect. 3.2). Furthermore, the results of the best machine systems submitted to PASCAL VOC's leaderboard are set in relation to the human agreement (Sect. 3.3). Finally, the results are made comparable in a systematic manner (Sect. 3.4).

3.1 Experimental Design

We have randomly selected 1,159 images from the PASCAL VOC test set. The relatively high number of images - compared to other studies - allows us to also obtain statistically relevant insights into human performance for less frequent concepts. For example, the concept cow is visible only in 34 out of 1,159 images, whereas the concept person occurs in 420 images. However, using PASCAL VOC's test set comes along with the disadvantage that the ground truth data are not available, in contrast to training and validation data. On the other hand, submission results are available only for the test set at PASCAL VOC's homepage. Therefore, we created ground truth for this test data subset by ourselves.

In total, twenty-three students (3rd and 4th year) were asked to annotate images of the test set with respect to 20 concept categories (see also Table 1), 18 students participated in the first and five other students participated in the second experiment. They were rewarded 25€ for participating in the experiment. All students were members of the Department of Electrical Engineering and Information Technology at the Jena University of Applied Sciences.

The participants were instructed to label images with respect to the presence of objects of 20 categories but without localizing them. Multiple object classes can be visible in an image, i.e., it is a multi-labeling task. This task corresponds to the classification task of the VOC challenge 2012. The study was further divided in two experiments. In the first experiment, the participants were instructed without using the PASCAL VOC annotation guidelines [1], since we aimed at measuring human performance based on common sense and existing knowledge about categories of daily life. In the second experiment, the participants were asked to annotate the images according to the PASCAL VOC guidelines.

The annotation process was divided in four batches that consisted of a slightly decreasing number of images. After each batch, the participants were allowed to make a break of 10–15 min. The annotation process had to be completed within four hours. The images were presented to all participants in the same order. They had to mark the correct object categories via corresponding checkboxes. When a user has finished annotating an image, he proceeded with the next image. The software did not allow users to return to a previously annotated image. All users completed the task within the given time limit.

3.2 Measuring Inter-Coder Agreement: Krippendorff's α

Krippendorff's α (K's α)[10,11] measures the agreement among annotators and is widely used in the social sciences to evaluate content analysis tasks. K's α is a generalization of several known reliability measures and has some desirable properties [11], it is (1) computable for more than two coders, (2) applicable to any level of measurement (ordinal, etc.) and any number of categories, (3) able to deal with incomplete and missing data, and (4) it is not affected by the number of units. In its general form K's α is equal to other agreement coefficients:

$$\alpha = 1 - \frac{D_o}{D_e}, \quad \text{where} \tag{1}$$

D_o is the observed disagreement and D_e is the expected disagreement due to chance. Krippendorff discusses differences of K's α with respect to other agreement coefficients [10] as well as explains its computation for various situations (depending on the number of coders, missing data, level of measurement, etc.) [11]. Hayes and Krippendorff provided a software that computes K's α [6].

3.3 Results for Inter-Coder Agreement

Agreement When *not* Using VOC Guidelines. The experimental results are displayed in Table 1 for the 1159 images of the PASCAL VOC test set. They

show the inter-rater agreement among the 18 coders by means of K's α. Across all concept categories and users, K's α is 0.913. The largest agreement among the annotators is observable for the three categories airplane, cat, and bird, whereas the categories dining table, chair, and potted plant yield the lowest agreement.

Table 1. User agreement (K's α) on a subset of PASCAL VOC test set, number of samples per category in this *subset*, and best machine-generated results (AI-1 and AI-2, avg. precision) on the *whole* test set.

Concept	#Samples per concept	Human (K's α)	AI-1	AI-2
Airplane	77	0.980	0.986	0.998
Cat	96	0.978	0.955	0.990
Bird	101	0.976	0.934	0.976
Dog	123	0.974	0.947	0.989
Sheep	35	0.970	0.874	0.950
Cow	34	0.960	0.821	0.943
Horse	42	0.959	0.929	0.985
Bus	45	0.956	0.910	0.959
Train	55	0.953	0.960	0.987
Boat	46	0.940	0.922	0.964
Motorbike	55	0.938	0.921	0.972
Bicycle	68	0.909	0.860	0.947
TV monitor	63	0.898	0.827	0.942
Person	420	0.895	0.950	0.988
Car	126	0.848	0.836	0.948
Sofa	79	0.796	0.678	0.868
Bottle	93	0.761	0.654	0.836
Dining table	61	0.737	0.796	0.881
Chair	123	0.716	0.734	0.904
Potted plant	59	0.668	0.594	0.768
Overall/MAP	–	0.913	0.854	0.940

Some interesting observations can be made. First, larger deviations of inter-coder agreement are observable for the different categories. Applying the rule of thumb that $\alpha > 0.8$ corresponds to a "reliable" content analysis [10], it turns out that the users' annotations cannot be considered as such for five categories: sofa, bottle, dining table, chair, and potted plant (K's α even only 0.67).

Table 1 shows also the results AI-1 and AI-2, which are the best submissions at PASCAL VOC's leaderboard website[1] for the competitions comp1 and comp2,

[1] http://host.robots.ox.ac.uk:8080/leaderboard/main_bootstrap.php.

respectively, by means of average precision (AP). The difference between the two is that comp2 is not restricted to the training set. Please note that these results of the user study and the VOC submissions are not directly comparable at this stage due to two reasons. First, the users annotated only a subset of the original test set. Second, different evaluation measures are used. In particular, the measures differ in the way how agreement by chance is considered. A more fair comparison resolving these issues is conducted in the subsequent section.

Anyway, when we set the inter-rater agreement in relation to the performance of the currently best machine results, we find that the correlation of K's α and AP with respect to the categories is 0.88 (AI-1) and 0.89 (AI-2), respectively. In particular, it is observable that the machine learning approaches perform worst for the same five categories as humans do.

Agreement When Using VOC Guidelines. In this experiment, it is investigated whether the inter-coder agreement is improved when more precise definitions are provided to the users. For this purpose, we have asked five other students (from the same department) to annotate the same 1159 images – this time based on the PASCAL VOC annotation guidelines [1]. The guidelines give some hints how the annotator should handle occlusion, transparency, etc.

Interestingly, the inter-coder agreement was not improved by using the guidelines. The inter-coder agreement in this experiment (measured again by K's α) is 0.904, in contrast to 0.913 without guidelines. The difference between the mean values of K's α with respect to all 20 categories is not statistically relevant (paired student's t-test, two-sided, significance level of 0.05), i.e., the reliability of human annotations is equal in both experiments. Although the participants used the annotation guidelines, the annotators did not achieve a better agreement in this experiment.

3.4 Comparing Human and Machine Performance

Since ground truth data for PASCAL VOC 2012 test set are not publicly available, we have created ground truth data for the related subset on our own. The ground truth data have been created according to PASCAL VOC annotation guidelines (see above). Critical examples were discussed by three group members (experts), two of them being authors of the paper. If a consistent agreement could not be achieved, then the example was removed for the related category.

In addition, we have trained a convolutional neural network (called AI-3 from now on) and evaluated its performance on the 1159 images. This allows us to apply AI-3 on the whole PASCAL test set as well as on the subset used in the human annotation study. Finally, this link enables us to compare human and machine performance for visual concept classification on PASCAL VOC test set data. We use the convolutional neural network of He et al. [7] consisting of 152 layers, which we fine-tuned on the PASCAL VOC training dataset. The network was originally trained on the ImageNet 2012 dataset [18]. Furthermore, we have reduced the number of output neurons to the number of classes (in this

case 20) and used a sigmoid transfer function to solve the multi-class labeling task of PASCAL VOC. Seven additional regions are cropped and evaluated per test image in the classification step to achieve better results. The mean average precision (MAP) for AI-3 is very similar for both datasets (0.871 vs. 0.867 on subset), although the difference of AP (for the subset and the whole test set) is larger for some classes, e.g., chair and train. The system AI-3 performs slightly better than AI-1. The results on both sets are depicted in Fig. 1.

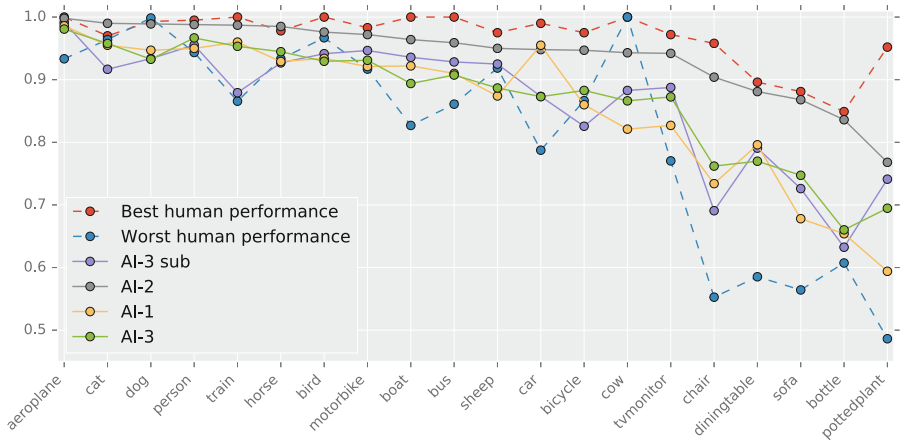

Fig. 1. Results (average precision in %) of the best and the worst human annotator, as well as of the best PASCAL VOC leaderboard submissions for comp1 and comp2.

Now, based on this analysis, the ground truth data allow us to compare human and machine performance using the same evaluation measure (average precision). But one issue still needs to be resolved. The annotators label the relevance of images only with "0" or "1" (in contrast to the real-values system scores). The (random) order of images labeled with "1" affects the measure of average precision. Due to this, we have calculated the "best case" (oracle) and the "worst case" possible retrieval results based on the human annotations.

Regarding the "best case" in the first experiment, there are three human annotators that are significantly better than AI-2 (paired student's t-test, one-sided, significance level of 0.05). These annotators achieve a mean average precision of 96.5 %, 96.3 %, and 95.3 %, respectively. However, regarding the "worst case" ordering, all human annotation results are significantly worse than AI-2, the best result (of the worst cases) achieves a mean average precision of 91.6 %. In other words, the results are sensitive to the ranking order of the images which are labeled as "relevant".

The results for the best annotator of our second experiment (using VOC guidelines) are similar. The best "best case" mean average precision is 96.9 %, this is the only human result of experiment 2 that is significantly better than

AI-2. Again, in case of "worst case" ordering, AI-2 is significantly better than all human annotation results.

Regarding both experiments, we have also estimated the best annotator (coming from experiment 2, with respect to best case ordering) and the worst annotator (coming from experiment 1, with respect to worst case ordering) with respect to our ground truth data. The results are displayed in Fig. 1. Again, the best results of the machine vision systems are presented as well. Regarding oracle results of the best human annotator, the automatic AI-2 system is only better for the two categories horse and cat. The AI-2 system achieved better results than AI-3. This can be explained by the fact that AI-3 relied only on the official training set, whereas AI-2 used additional data. However, the AI-3 system outperforms the worst human annotator in the most cases.

Overall, the results indicate that the automatic system AI-2 indeed reaches performance comparable or even superior to humans. To be more precise, even when (artificially) optimizing the ranking of the human annotations with respect to ground truth data, the system AI-2 still was better than 9 participants and was on a par with 6 participants (only 3 human "best case" results were better than AI-2) in experiment 1 (all results based on a paired student's t-test, one-sided, significance level of 0.05). It is similar for experiment 2: AI-2 is better than 2 human "best case" results (both experiments: 11), on a par with another 2 results (both experiments: 8), and a single human annotator performed better (both experiments: 4). In other words, the system AI-2 is at least on a par with 83 % (or better than 48 %) of the human participants in our study.

4 Measuring Machine Performance at Human Level

4.1 Issues of Measuring Human-Level Performance

The performance of multimedia retrieval systems is often measured by average precision (AP). However, this measure has some known drawbacks. First, the measure depends on the frequency of a concept. Considering the definition of average precision, it is clear that the lower bound is not zero but determined by the frequency of relevant documents in the collection. When randomly retrieving documents, the average precision will be equal to a concept's frequency on average. This has been stressed, e.g., by Yang and Hauptmann who suggested the measure ΔAP (delta average precision) to address this issue [24]. Second, the upper bound of 1.0 is also not reasonable. If we consider that the agreement in our user study is below 0.8 for five categories and is even only 0.67 for potted plant, the question arises how we have to interpret an average precision of, for example 67 % and 100 %, respectively. If two raters agree with K's α only with 0.67, and one rater corresponds to ground truth, it is basically possible that the 67 %-result has the same quality as the 100 %-result (from another perspective, in another context). Hence, the question remains: how can we formally measure whether a machine-based result is comparable to or even better than a human result? This question is addressed in the subsequent section.

4.2 An Experimental Methodology and Two Indices for Comparisons with Human-Level Performance

In this section, we propose an experimental methodology and two novel easy-to-use measures, called human-level performance index ($HLPI$) and human-level performance ranking index ($HLPRI$). The proposed methodology is aimed at providing a systematic guidance to measure human-level performance in multimedia retrieval tasks. It is assumed that a visual or auditory concept is either present or not in a multimedia document. Furthermore, it is assumed that a standard benchmark dataset is available. First, ground truth data G should be created by knowledgeable experts E of the related domain. If possible, the reliability of these expert annotations should be measured as well (K's α should exceed at least 0.8 as a rule of thumb) in order to ensure that the relevance of categories is well-defined. Critical examples should be subsequently discussed among the group of experts E. If no consistent decision is possible, then such examples should be appropriately marked or discarded from the dataset.

The group of human participants H should consist of at least five annotators/coders, who share a similar knowledge level regarding the target domain (e.g., experts, if performance is to be measured at expert level). The annotations of the group H are used to evaluate human performance in the given task. The annotation process should be conducted in a well-defined setting. The latter two criteria (same knowledge level, well-defined setting) normally preclude a crowdsourcing approach. Apart from other measures, the inter-rater agreement among the annotators should be also estimated by Krippendorff's α. K's α is suggested since it can potentially also deal with other levels of measurement (than binary). Then, the agreement (accuracy) a_{human} is the median of the agreement scores (e.g., measured by K's α or AP) of the coders H with respect to G. The machine-generated result is also compared to G, yielding the agreement a_{machine}. Often, several instances of a machine system, e.g., caused by different parametrizations or training data, exist. In this case, it is also reasonable to test all these instances with respect to G and use the median as a_{machine}. In this way, it can be prevented that a system is better only due to fine-tuning or by chance. Then, the human-level performance index is defined as

$$HLPI = \frac{a_{\text{machine}}}{a_{\text{human}}}, \tag{2}$$

assuming that human inter-coder agreement is better than chance, i.e., $A_{\text{human}} > 0$. If $HLPI > 1.0$, then machine performance is possibly better. However, this has to be verified and ensured by an appropriate statistical significance test.

A second measure the human-level performance ranking index ($HLPRI$) is suggested. This index is based on a sorted list in descending order according to the agreement measurements of the n human annotators with respect to ground truth data G. Let b be the number of machine results to be evaluated that are

better than human annotation results, and let w be the number of machine annotation results that are worse (in the sense of statistical relevance). Then, the human-level performance ranking index is defined by

$$HLPRI = \frac{b+1}{w+1} \tag{3}$$

$HLPRI$ of AI-2 is 2.4 and 1.5 in our experiments 1 and 2, respectively.

5 Conclusions

In this paper, we have investigated the question whether today's automatic indexing systems can achieve human-level performance in multimedia retrieval applications. First, we have presented a brief survey comparing human and machine performance in a number of visual and auditory recognition tasks. The survey has been complemented by two extensive user studies which investigated human performance in an image annotation task with respect to a realistic photo collection with 20 common categories of daily life. For this purpose, the well-knwon PASCAL VOC benchmark has been used. We have measured the human inter-coder agreement by Krippendorff's α and observed that the reliability of annotations noticeably varies for the concepts. Krippendorff's α was below 0.8 for 5 out of 20 categories, which indicates that these categories are not well-defined and are prone to inconsistent annotation. This is an issue for the creation of ground truth data and subsequent evaluation as well.

In addition, we have carefully compared human and machine performance for image annotation. It turned out that the best submission at PASCAL VOC's leaderboard is better than 11 or at least on a par with 19 out of 23 participants of our study. This indicates that the submission has indeed reached above-average human-level performance for the annotation of the considered visual concepts.

We have also addressed the issue of measuring human-level performance of multimedia analysis and retrieval systems in general. For this purpose, we have suggested an experimental methodology that integrates the assessment of human-level performance in a well-defined manner. Finally, we have derived two easy-to-use indices for measuring and differentiating human-level performance.

References

1. VOC2011 annotation guidelines. http://host.robots.ox.ac.uk/pascal/VOC/voc2011/guidelines.html. Accessed 29 Mar 2016
2. Bailer, W., Schallauer, P., Thallinger, G.: Joanneum research at TRECVID 2005-camera motion detection. In: Proceedings of TRECVID Workshop, pp. 182–189 (2005)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. IEEE (2009)

4. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
5. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
6. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. **1**(1), 77–89 (2007)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint (2015). arXiv:1512.03385
8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 1026–1034 (2015)
9. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In: Proceedings of the International Conference on Multimedia Retrieval (ICMR), p. 29. ACM (2011)
10. Krippendorff, K.: Reliability in content analysis: Some common misconceptions and recommendations. Hum. Commun. Res. **30**, 411–433 (2004)
11. Krippendorff, K.: Computing Krippendorff's alpha-reliability (2011). http://repository.upenn.edu/asc_papers/43/. Accessed 29 Mar 2016
12. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: International Conference on Computer Vision (ICCV), pp. 365–372. IEEE (2009)
13. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
14. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). doi:10.1007/978-3-319-10602-1_48
15. Nowak, S.: Evaluation methodologies for visual information retrieval and annotation. Ph.D. thesis (2011)
16. Parikh, D., Zitnick, C.L.: The role of features, algorithms and data in visual recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2328–2335. IEEE (2010)
17. Phillips, P.J., Scruggs, W.T., O'Toole, A.J., Flynn, P.J., Bowyer, K.W., Schott, C.L., Sharpe, M.: FRVT 2006 and ICE 2006 large-scale results. Natl. Inst. Stand. Technol. NISTIR **7408**, 1 (2007)
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. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
19. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. IEEE (2015)
20. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708. IEEE (2014)
21. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Semantic annotation and retrieval of music and sound effects. Trans. Audio Speech Lang. Process. **16**(2), 467–476 (2008)

22. Weyand, T., Kostrikov, I., Philbin, J.: PlaNet-photo geolocation with convolutional neural networks. arXiv preprint (2016). arXiv:1602.05314
23. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492. IEEE (2010)
24. Yang, J., Hauptmann, A.G.: (Un)reliability of video concept detection. In: Proceedings of the International Conference on Content-based Image and Video Retrieval (CIVR), pp. 85–94. ACM (2008)
25. Zahálka, J., Rudinac, S., Worring, M.: Analytic quality: evaluation of performance and insight in multimedia collection analysis. In: Proceedings of the International Conference on Multimedia (MM), pp. 231–240. ACM (2015)