This CVPR 2021 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Less is More: Pursuing the Visual Turing Test with the Kuleshov Effect

Gustavo Olague
CICESE Research Center
olague@cicese.mx

Matthieu Olague
Anahuac University Queretaro
matthieu.olague03@anahuac.mx

Angel R. Jacobo-Lopez
CICESE Research Center
arjacobo@cicese.edu.mx

Gerardo Ibarra-Vazquez
Autonomous University of San Luis Potosi
gerardo.ibarra.v@gmail.com

Abstract

The Turing test centers on the idea that if a computer could trick a human into believing that it was human, then the machine was deemed to be intelligent or indistinguishable from people. Designing a visual Turing test involves recognizing objects and their relationships on images and creating a method to derive new concepts from the visual information. Until now, the proposed visual tests heavily use natural language processing to conduct the questionnaire or storytelling. We deviate from the mainstream, and we propose to reframe the visual Turing test through the Kuleshov effect to avoid written or spoken language. The idea resides on elucidating a method that creates the concept of montage synthetically. Like the first days of cinema, we would like to convey messages with the interpretation of image shots that a machine could decipher while comparing it with those scored by humans. The first implementation of this new test uses images from a psychology study where the circumplex model is applied to rate each image. We consider five deep learning methodologies and eight optimizers, and through semiotics, we derive an emotional state in the computer. The results are promising since we confirm that this version of the visual Turing test is challenging as a new research avenue.

1. Introduction

Imagine we walk into the house, and our companion robot is reviewing our oldest son's final-year college homework on the Kuleshov effect. This might sound like a black mirror tale whose futuristic scenario is common among Sci-Fi and fantasy lovers. However, to understand the implications, we need to come back to the first days of cinema. In the Soviet Union, a revolution on silent film was developed by film theorist Lev Kuleshov, the father of Soviet montage theory. In those days, silent movies needed to convey messages without the written word. We hold to Kuleshov's viewpoint that cinema's essence is editing, juxtaposing one shot with another, and introducing the idea in computer vision that image understanding could be put to the test by attempting to recreate a mental phenomenon known as the Kuleshov effect in silico.

1.1. Related Work

The Turing test, originally called the imitation game by Alan Turing in 1950, is a test of a machine's ability to exhibit intelligent behavior equivalent to or indistinguishable from a human [33]. Turing's idea consists of a natural language conversation among a human evaluator, another human, and a machine designed to generate human-like responses. The evaluator is aware that one of the two partners is a machine, and all participants would be separated from one another. If the evaluator cannot reliably tell the machine from the human, the machine is said to have passed the test. The test results do not depend on the machine's ability to give correct answers to questions, only how closely its answers resemble those a human would give. There are numerous drawbacks as a rigorous and practical means of assessing progress toward human-level intelligence like language focus, complex evaluation, subjective evaluation, and difficulty measuring incremental progress [1]. Nevertheless, the test is considered the holy grail of computer intelligence. To reach the ultimate goal of machine intelligence, where machines supplement humans, the research community needs to resolve several hurdles. Some challenges may be tackled purely technologically; others require insights from sociology and psychology to break new ground [3].

Thanks to the advances in machine learning through a neural network technique called deep learning, the research community has a special interest in what is known as the visual Turing test. A first attempt consists of designing a written test that uses binary questions to probe a system's ability to identify attributes and relationships in addition to recognizing objects [9]. The problem of visual question answering (VQA) derives from the visual Turing test proposed by Geman *et al.* in 2016. However, the world's inherent structure and language bias tend to be a simpler signal for learning than visual modalities, resulting in VQA models that ignore visual information leading to an inflated sense of their capability [11] with no clear answer to this dilemma. Moreover, despite great advances in multiple areas (object recognition, speech recognition, board games, video games, and control) deep neural networks (DNN) differ from human intelligence in crucial ways. Also, the success of DNN came with several major problems (adversarial attacks, high-computational cost, high amount of labeled data, lack of invariance, spatial relationship, and explainability) that can hamper their applicability.

According to [21] progress in cognitive science suggests that truly human-like learning and thinking machines require new approaches that define what and how machines learn. They argue that such method builds causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems, ground learning in intuitive theories of physics and psychology to support and enrich the knowledge that is learned, and harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. Compositionality is the classic idea that new representations can be constructed through the combination of primitive elements. In computer programming, primitive functions can be combined to create new functions, and these new functions can be further combined to create even more complex functions. This function hierarchy provides an efficient description of higher-level functions, such as a hierarchy of parts for describing complex objects or scenes [25]. Structural description models represent visual concepts as compositions of parts and relations, which provides a strong inductive bias for constructing models of new concepts. Lake *et al.* describe a computational model that learns in a similar fashion and does so better than current deep learning algorithms. The model classifies, parses, and recreates handwritten characters, and can generate new letters of the alphabet that look right as judged by Turing-like tests of the model's output in comparison to what real humans produce [22].

Emotion classification, how one may distinguish or contrast one emotion from another, is a contested issue in emotion research and affective science. Researchers have approached the classification of emotions from two fundamental viewpoints: 1) that emotions are discrete and fundamentally different constructs, 2) that emotions can be characterized on a dimensional basis in groupings. Recently, researchers study psychological theories of emotion to develop an intelligent system of decision-making for autonomous and robotic units focusing on verbal and non-verbal agent communication [18]. In [6] also studied limited Turing test for social-emotional intelligence with a video game-like virtual environment. In recent years, with the increasing use of digital photography, emotional semantic image retrieval appears as an appealing subject of study. In [23] authors study affective image classification using features inspired by psychology and art theory. From a psychology standpoint, affective pictures are widely used in studies of human emotions [29]. The objects or scenes shown in affective pictures play a pivotal role in eliciting particular emotions. Affective processing derives from local and global image properties as well as image composition. A survey about affective image content analysis is provided in [35]. Today, there is a gap between image content and emotional response, and some researchers attempt to close it using high-level concepts [2]. In the article, the authors explain how to simplify the problem through databases' division to understand affective classification. In [16] authors attempt to infer communicative intents of images as a visual persuasion. Also, they strive to elucidate this ability from a single image. Authors consider thousands of different types of emotions or feelings that can attribute to the image's main target. They select ten emotional traits, personality traits and values (six additional dimensions), and overall favorability (a binary variable reflecting positive or negative status). Also, they consider syntactical attributes like facial display, body cues-gestures, and scene context. Images and human emotions were also investigated through a mixed bag of emotions approach [26]. The aim is to model, predict, and transfer emotion distributions. They study the psychological problem of identifying the primary emotions in an image and manipulating the image to evoke a different response by adjusting color tone and texture-related features. Finally, we would like to mention a work where affective understanding in the film plays an essential role in sophisticated movie analysis, ranking, and indexing [13]. They recognize the lack of work to close the gap between low-level features and emotion. Authors follow a systematic approach grounded upon psychology and cinematography to address several critical issues in affective understanding. They follow a probabilistic method identifying a set of categories and steps for classification. Also, they use many of the concepts that we have described in the reviewed work and recognize that this subject remains a largely unexplored field.

1.2. Problem Statement

The reviewed literature exposes that current engineering trends demand new approaches based on psychology, physics, and symbolic learning to answer the question Can machines think? following the imitation game. In [7] the author suggests that for such a move to be successful, the test needs to be relevant, expansive, solvable by exemplars, unpredictable, and lead to actionable research. Although the test is at the top of artificial intelligence, its reliance on language, whilst insightful for partially solving the problem, has put progress on the wrong foot, prescribing a top-down approach for building thinking machines. Instead of this path, Crosby proposes a bottom-up approach founded in animal cognition tests. We propose to avoid the problem of language–hence the top-down *vs.* bottom-up dilemma–following the original pathway of cinema made during the silent film era by adapting the Kuleshov theory to the task of image understanding in computer vision. Unlike previous works that have studied the Turing test or the imitation game, the goal is not to develop a new method that necessarily outperforms previous methods. In other words, we do not intend to make a technological contribution but to reveal a scientific avenue that can serve in the long run to design a visual Turing test. Here, we provide the first method to illustrate the idea.

Figure 1 illustrates the Kuleshov effect using images from the datasets described in Section 2.1. In the original experiment, Kuleshov and his protégé Pudovkin edited a short film in which a shot of an expressionless face actor (Ivan Mosjoukine) followed by either a bowl of soup, a dead woman in a coffin, or a little girl playing affects the perception. Pudovkin said that the audience viewing the sequences reported three different judgments of Mosjoukine's facial expression: heavy pensiveness, deep sorrow, and happiness, respectively. Although the shot of the actor's neutral face was identical in all three scenarios, the context provided by the subsequent film shot affected the audience's interpretation of the actor's emotion conveyed by his facial expression [28]. The effect of visual context on the interpretation of facial expression from an actor's face was successfully tested using isolated photographic stills instead of the typical dynamic film sequences used to demonstrate the Kuleshov effect [24].

The Kuleshov effect in the machine attempts to simulate a psychological phenomenon where a neutral face followed by an image evoking two affective dimensions produce through montage–we need a technique to emulate such phenomenon–the idea of a facial expression. For such an idea (montage) to be implemented on the computer, we need to introduce some concepts borrowed from semiotics [5, 4]. Note that montage is different from compositionality since the former creates a novel idea not present in the primitive elements. Semiotics is the study of sign processes (semiosis), any activity, conduct, or process that involves signs. A sign is defined as anything that communicates a meaning that is not the sign itself to the sign's interpreter.

Semiotics deals with the relationships that arise when something represents something else. This is the signifier-signified relationship. In film semiotics, Christian Metz emphasizes that multiple potential sequences give structure to a film event whose signified can then be variable. The signified is the idea of the object, for example, a cat. The signifier can be an icon (image cat), an index (image of cat's footprint), or a symbol (Chinese cat symbol). To achieve the Kuleshov effect, we have that this consists of the combination of (classically) two meanings (signified) for the extraction of a third whose meaning (signifier) is not necessarily represented in a visual scene. This effect is widely used as part of the montage technique in cinematography. Here, we attempt to emulate such effect through machine interpretation of visual stimuli to produce an output not present in the original sequence, see Figure 1.

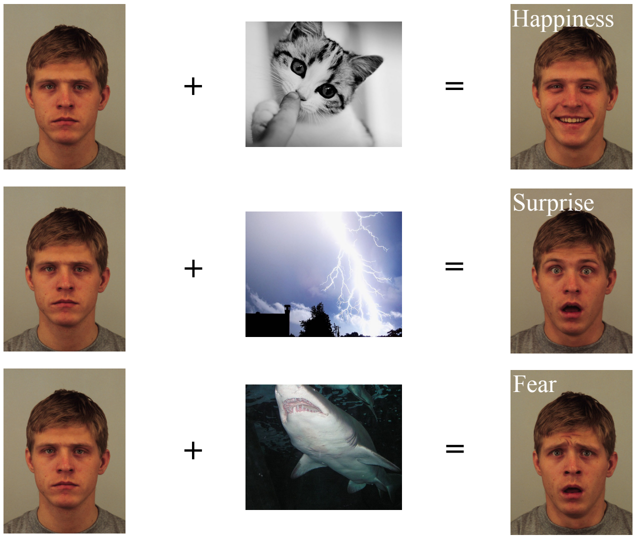

Figure 1. The Kuleshov effect in the machine attempts to simulate a psychological phenomena where a neutral face followed by an image evoking two affective dimensions produce through montage (emulate phenomenon) the idea of a facial expression.

1.3. Contributions

In this work, we propose a new way of pursuing the visual Turing test. The idea consists of replicating the phenomenon of montage in the computer. An immediate benefit is to avoid the usage of natural language processing while pursuing the goal of inquiring about the logic of a machine vision system regarding computer's interpretations of images. This perspective allows us to frame the Turing test within a purely visual processing scenario where artificial cognitive aspects are confronted using several meaningful images while altering the order in a sequence to produce a new concept not explicit in the original shot.

2. Methodology

The goal of this section is to derive a visual Turing test that doesn't rely on language. Next, we describe the circumplex model and the flowchart of the proposed system.

2.1. The Circumplex Model of Affect

The circumplex model of emotion, also known as the circumplex model of affect, can describe the complexity of emotions and their representation [30]. Researchers use the circumplex concept to study how different emotions relate using a circular depiction of multiple variables' similarities. Researchers studying emotion focus on specific core affects and create a circumplex representation of them, with variables having opposite values or characteristics (*i.e.* delighted-miserable, happy-sad, relaxed-distressed) displayed at opposing points on the circumplex. In contrast, variables having highly similar characteristics are displayed adjacent to one another on the circumplex. In other words, the similarity (and correlation) between elements declines as the distance between them on the circle increases. Hence, a circumplex model is a model-type that shows unique relationships within a visual framework. It is a circular model divided into quadrants with axes of crossed continua. Thus, the circumplex model of affect suggests that emotions distributed in a two-dimensional circular space containing arousal and valence dimensions emerged from environmental stimuli. Arousal represents the vertical axis, and valence represents the horizontal axis, see Figure 2. The circumplex is a dimensional model whose goal is to conceptualize human emotions by defining where they lie in two or three dimensions. The theory states that an ordinary and interconnected neurophysiological system is responsible for all affective states. The idea in this work is to relate the KDEF database's basic emotions with the OASIS database's visual stimuli following the Kuleshov effect.

2.2. Visual Turing Test Flowchart

The test consists of simplifying the imitation game since we first want to verify the capacity of current deep learning technologies. Inspired by Kuleshov's ideas, we juxtapose different concepts (*i.e.*, images of a city, money, fire, and so on) with a neutral face through artificial montage, resulting in an evoked facial expression not seen initially in any of the signifiers. We use information collected by psychologists as ground truth from the Open Affective Standardized Image Set (OASIS) study, where human beings respond to a wide variety of visual stimuli [20]. Emotions are measured in OASIS using normative ratings on two affective dimensions (valence and arousal), which are used to abstract both the physical and social worlds on images. Psychologists recruited a diverse set of participants to gauge their affective responses to the images. Valence is the degree of positive or negative affective response that the image evokes, and arousal determines the intensity of the affective response that the image evokes. The ratings obtained covered much of the circumplex space and were highly reliable and consistent across gender groups. With only a few thousand pixels, images can depict an unlimited array of people, objects, and scenes. They can evoke a range of affective responses, which we simplify by superposing the results of the two affective dimensions into seven facial expressions (neutral, happy, angry, afraid, disgusted, sad, and surprised) obtained from the Karolinska Directed Emotional Faces (KDEF) dataset [10]. We select a few subjects with a front profile from the dataset to demonstrate the kuleshov effect.

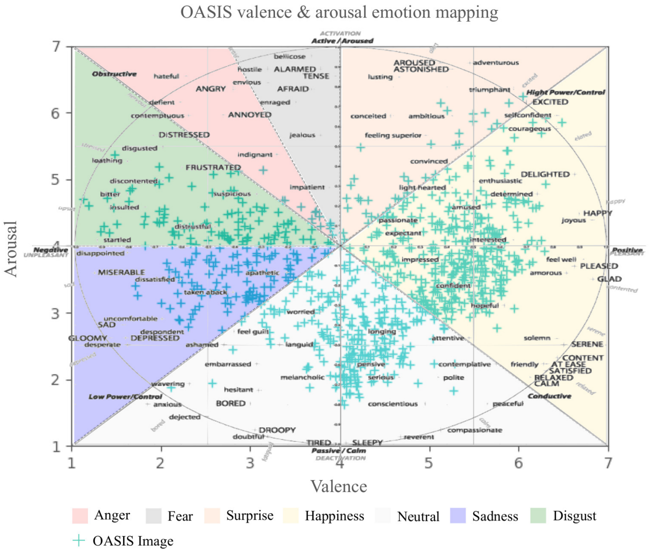

Figure 2. Valence and arousal ratings for OASIS images as portrayed by human beings. Values are classified using the circumplex model of affect, where each colored region in the map shows an emotion obtained from the KDEF database.

Figure 3 illustrates two critical aspects of our proposal: 1) how we train CNN models based on the OASIS normative ratings, and 2) how we use the circumplex model to map emotions. The idea put forward of our visual Turing test consists of learning the mappings between the input images and the scored values of valence and arousal given by people with several machine learning methods, see Table 2.1. We built a sequence by juxtaposing a subject's neutral face with a selected image from OASIS. The idea is to identify the subject in the neutral face, and with the computed scores achieved by the mapping, the system could render the appropriate facial expression of the related subject, which matches the evoke displayed by people. The process of mapping emotions to familiar facial expressions allows us to establish common comparison grounds between humans and machines.

Note that the proposed methodology has a limit given by the number of facial expressions that humans can portray, since it becomes facially challenging to show differences between being happily amused, happily impressed, or hap-

CNN

AlexNet [19]	Replacement of final 1000-way softmax layer with two fully connected (FC) output neurons activated by a linear activation function.
DenseNet-121 [14]	Stacked a two neuron FC linear layer to perform a regression on valence and arousal values.
ResNet-18 [12]	Stacked a two neuron FC linear layer to perform a regression on valence and arousal values.
VGG-11(BN) [31]	Stacked a two neuron FC linear layer to perform a regression on valence and arousal values.
SqueezeNet [15]	Stacked a two-dimensional convolutional layer with 512 input channels and two output channels, with kernel and stride sizes of 1.

Optimizers

SGD [32]	learning rate $\alpha = 1 \times 10^{-5}$ momentum coefficient $\gamma = 0.9$
Adam [17]	learning rate $\alpha = 1 \times 10^{-4}$ weight decay norm $L_2 = 1 \times 10^{-5}$ exponential decay rate $\beta_1 = 0.9$ exponential decay rate $\beta_2 = 0.999$
AdaMax [17]	learning rate $\alpha = 1 \times 10^{-4}$ weight decay norm $L_2 = 1 \times 10^{-5}$ exponential decay rate β_1= 0.9 exponential decay rate β_2= 0.999 constant $\epsilon = 1 \times 10^{-8}$
AdaGrad [8]	learning rate $\alpha = 1 \times 10^{-4}$ weight decay norm $L_2 = 1 \times 10^{-5}$ constant $\epsilon = 1 \times 10^{-10}$
ADADELTA [34]	learning rate $\alpha = 1 \times 10^{-4}$ weight decay norm $L_2 = 1 \times 10^{-5}$ constant $\epsilon = 1 \times 10^{-10}$ constant $\rho = 0.9$
ASGD [27]	learning rate $\alpha = 1 \times 10^{-4}$ weight decay norm $L_2 = 1 \times 10^{-5}$ decay term $\lambda = 1 \times 10^{-4}$ smoothing constant $\eta = 0.75$ averaging starting point $t_0 = 1 \times 10^{6}$
RMSprop	learning rate $\alpha = 1 \times 10^{-4}$ momentum coefficient $\gamma = 0.9$ smoothing constant $\eta = 0.99$ constant $\epsilon = 1 \times 10^{-8}$
Rprop	learning rate $\alpha = 1 \times 10^{-5}$ etaminus set to 0.5 etaplis set to 1.2 step sizes from 1×10^{-6} up to 50

Table 1. This table shows the design of frameworks divided into two parts: on the left, we show the list of CNNs with their main modifications, and on the right, we present several optimizers tested on the experiments.

pily expectant, to name a few examples shown in Figure 5. In this way, we define a set of regions that classify the primary emotions by dividing the circumplex model into sixteen equal portions with seven bounds set along the radius of a circumference as seen in Figure 2. We assign the resulting seven slices to the KDEF facial expressions. Also, we evaluate the emotional response elicited by the OASIS image and valence-arousal values with the absolute distance between predictions and neighboring circumplex emotions calculated as a probability with the softmax function as follows:

$$\sigma(\overrightarrow{\mathrm{d}})_i = \frac{e^{d_i^{-1}}}{(\sum_{j=1}^{k} e^{d_j^{-1}})}, \tag{1}$$

where $\overrightarrow{\mathrm{d}} \in \mathbb{R}^k$ is an array comprising the measured Euclidean distances. Note that to attribute higher probabilities to nearest emotions, we used the inverse relation for d_i and d_j and integrated it into the formula.

3. Experimental Results

We implemented, trained, and tested a total of 40 frameworks into the previously delineated system workflow, making use of CNNs and optimizers included in Table 2.1. Figure 4 illustrates the optimization algorithm that resulted in the highest test accuracy for each network; these being AlexNet with Rprop, DenseNet-121 with Adam, ResNet-18 with SGD, SqueezeNet with AdaMax, and batch normalized VGG-11 with SGD. Furthermore, columns inside the figure provide insight for said individuals by using various graphics displayed along rows. The first row presents a set of scatter plots that contain valence and arousal values predicted by the machine during training. The similarity between frameworks is noticeable and comparable to the emotional responses elicited on human beings, an ob-

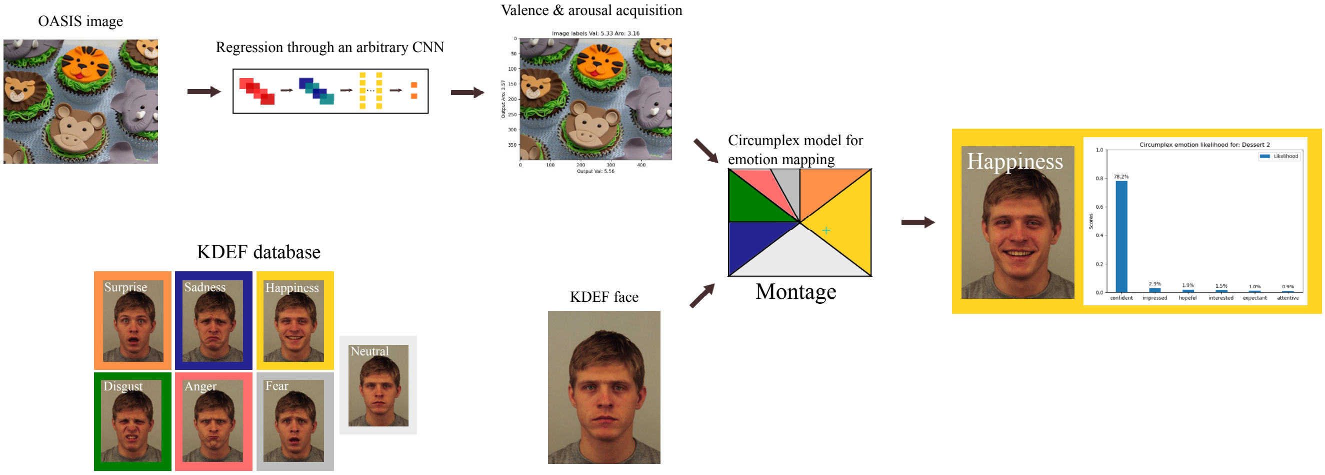

Figure 3. This diagram shows the proposal to emulate the Kuleshov effect, where the upper section shows how we can determine through arbitrary CNNs valence and arousal values from a given image taken from the OASIS database. The bottom part shows a neutral face taken from the KDEF dataset superposed with results of the two affective dimensions to generate one of seven facial expressions.

servation that we can understand through the visualization of adjacent confusion matrices. Such graphs aim to compare the machine predicted values with the average human emotions reported by psychologists on the OASIS study; thus, by observing the confusion matrix's columns, the corresponding network's prediction can be visualized. Analogously, the rows that comprise the confusion matrix show the ground truth–mean human response–that corresponds to the utilized photographs, and in the intersection of both emotional replies, we establish a comparison. Two valuable metrics were extrapolated from the achieved results to consolidate this comparison: the weighted accuracy and F1 score across classes.

As mentioned above, all these factors lead to the conclusion that during training, all convolutional neural networks were capable of reproducing with high fidelity the answers collected in the OASIS study. The few misclassified images are close to the correct subclass (emotion) according to the circumplex model. In the case of testing, a consistent drop in accuracy occurs as shown in the third and fourth rows of figure 4. Despite the efforts to minimize overfitting, the high variability in concepts presented a challenge to achieve correct emotional classification (*e.g.*, it is not equivalent to interchange happiness with neutral and happiness with disgust). The fifth row presents the evolution of the mean squared error across epochs, exemplifying the pattern memorization of deep learning frameworks.

4. Discussion

Figure 5 shows an underlying and fundamental problem causing emotional misclassification in CNNs, which goes beyond overfitting issues. While a human being reacts according to context and understands a presented scenario, architectures such as AlexNet react according to previously identified patterns. In the first part, we observe how two images addressing the same theme (fire) can have two opposite emotional reactions. On the left side, we appreciate how humans rightly agree to the visual stimulus, while on the right side, the machine erroneously comprehends the fire theme. In the last two rows, we observe results obtained with four different images (city, money, couple, and woman) considering human Vs. machine. Again it is remarkable the automatic pattern identification that contrasts with human perception.

The proposed experiments show how image understanding is an issue that these technologies cannot yet fully address. Moreover, the capability of memorization presented during training is preoccupying. Cases such as the ones discussed previously, taken from the set of images reserved for testing, would have been correctly classified if assigned to the training set as shown by the results in Figure 4. The proposed machine learning technique implies that if a more extensive database had been used, with more images and cases such as those previously presented, the various architectures would have had the opportunity to retain and identify a more significant number of patterns. Extending the database, in turn, does not solve the underlying problems. It only tries to cover them, where overfitting of CNNs manifests as a symptom of this stumbling block.

Other areas and cases of application for CNN methodologies are not exempt from these issues (memorization, size of database, image understanding), and as such, must be taken into account and consideration. Learning leaves an open road for research in identifying alternate solutions and methodologies that could prove helpful in fields of study, including–but not limited to–computer vision, pattern

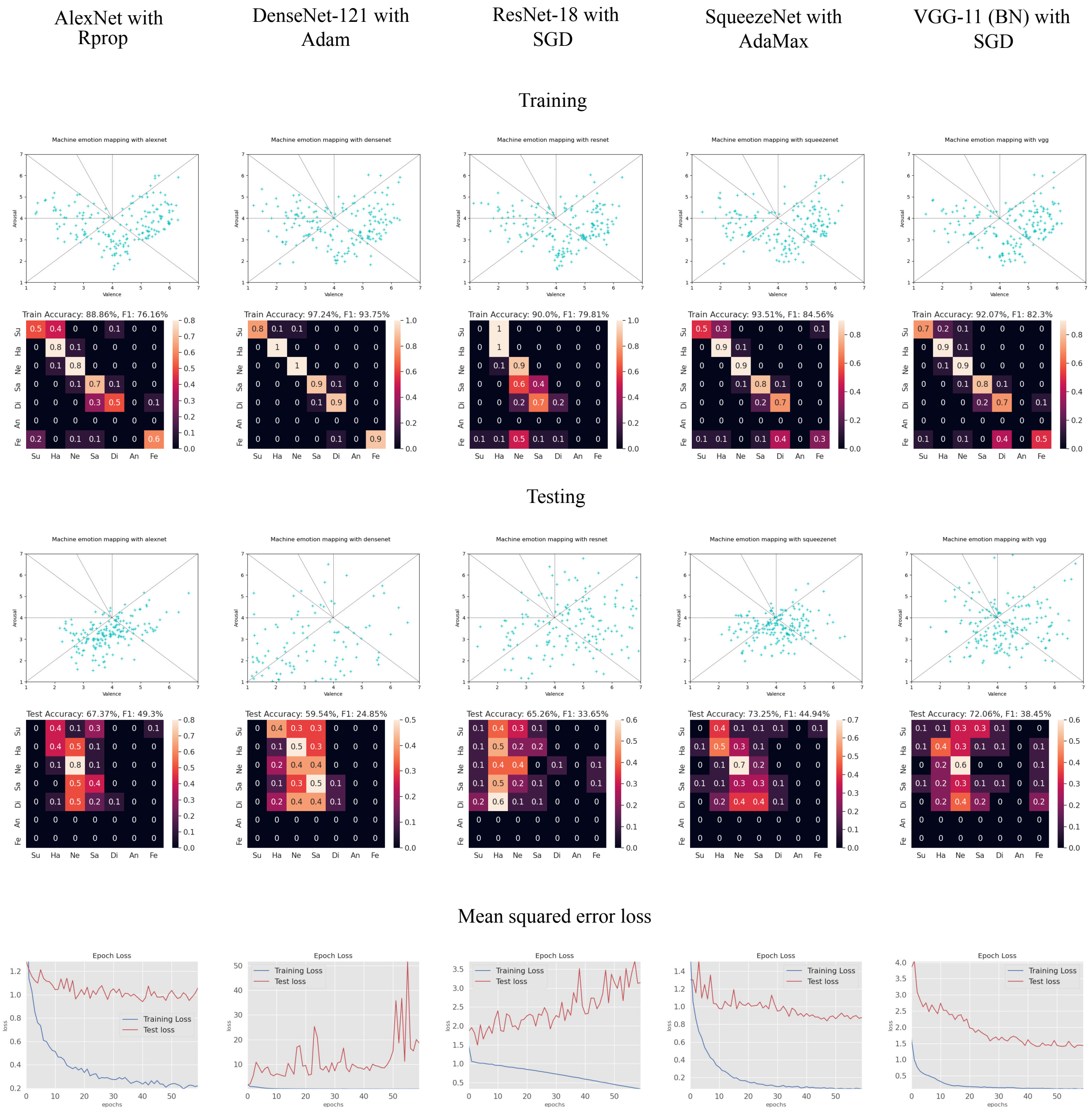

Figure 4. Emotion mapping of OASIS images using various convolutional neural networks (Alexnet, Densenet, Resnet, Squeezenet and VGG). Figures depict the optimizer (SGD, Adam, Adadelta, Adagrad, Adamax, ASGD, RMSprop, and Rprop), that yields best test results for the corresponding CNN. Additionally, difference between training and testing is displayed through scatter plots for the predicted valence and arousal values, confusion matrices that compare human vs. machine emotion classification, and the evolution of training and testing loss across epochs.

recognition, psychology, or even cinematography. On a final note, it is worth mentioning that applications in the film or video-game industry are possible for the current results, which are environments with known and static scenarios, where one of the tested frameworks could provide reliable and accurate results through training. The idea of extending the visual Turing test to more complicated videos through the Kuleshov effect is a challenging research area.

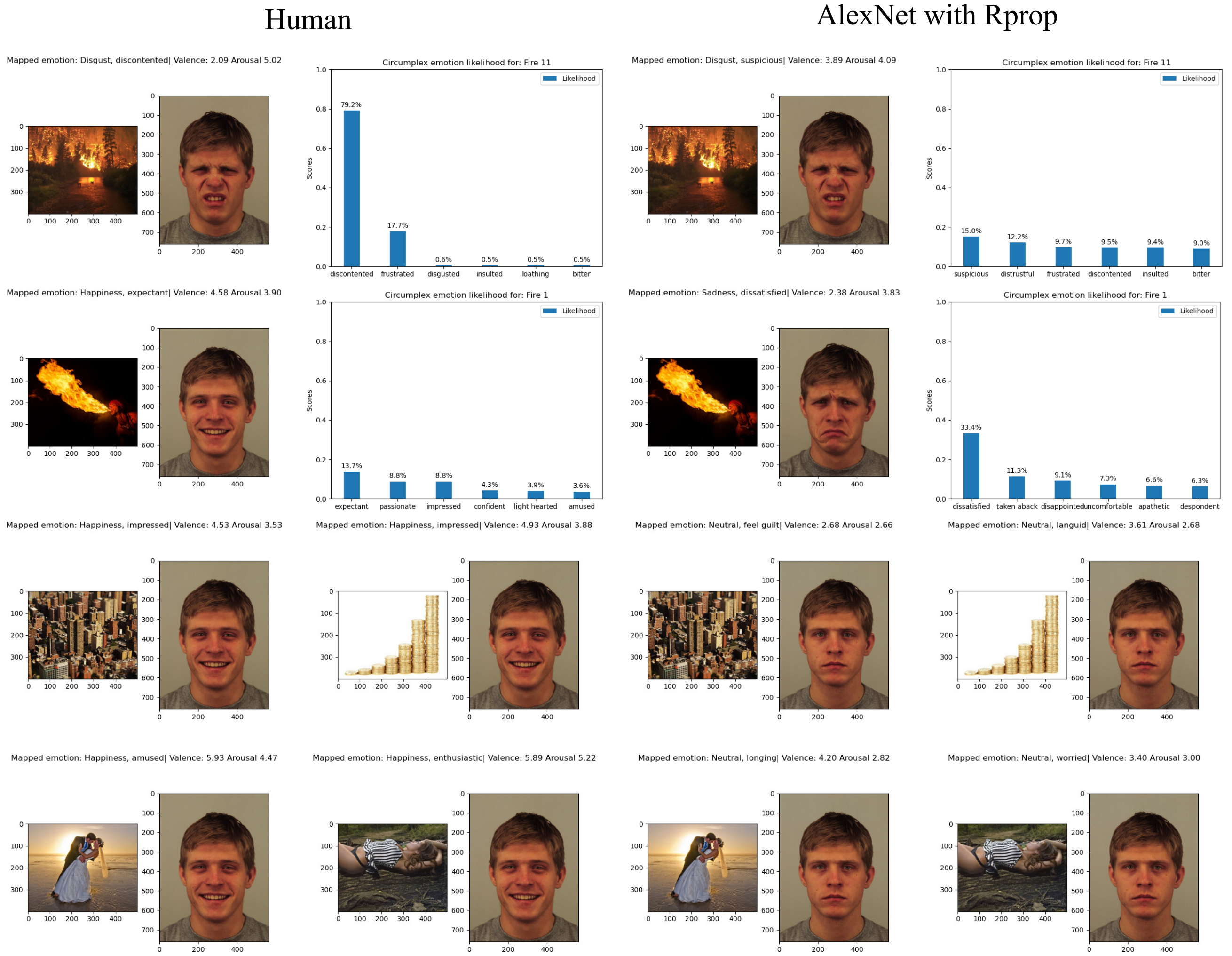

Figure 5. The following photographs show the importance of image understanding for accurate emotional interpretation. On the left, we can observe a human being's response towards visual stimuli taken from the OASIS database. On the right, we can appreciate AlexNet's reaction towards the same set of pictures, showing an expected–but incorrect–consistency due to the involved high-level knowledge.

5. Conclusion and Future Work

We devise a new framework for the visual Turing test through the Kuleshov effect. The conceptual design avoids using natural language processing based on the idea of synthetically creating montage from two different shots in a sequence of images. Like in the silent film era, we would like to convey messages by juxtaposing independent photographs. The proposed technique applies a test designed for humans to rate image content, and the results achieved with the computer reveals the difficulty of solving this task. In the future, we would like to test other deep learning approaches and take a step backward using computer vision and computational intelligence methods since we hold to the idea that less is more.

References

[1] Sam S. Adams, Guruduth Banavar, and Murray Campbell. I-athlon: Towards a multidimensional turing test. *AI Magazine*, 37(1):78–84, Apr. 2016.

[2] A. Ali, U. Shahid, M. Ali, and J. Ho. High-level concepts for affective understanding of images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 679–687, Los Alamitos, CA, USA, mar 2017. IEEE Computer Society.

[3] Bo Begole. Responsive media: media experiences in the age of thinking machines. *APSIPA Transactions on Signal and Information Processing*, 6:e4, 2017.

[4] Warren Buckland. *The Cognitive Semiotics of Film*. Cambridge University Press, 2000.

[5] Daniel Chandler. *Semiotics: The Basics*. Routledge, 3rd edition, 2017.

[6] Arthur Chubarov and Daniil Azarnov. Modeling behavior of virtual actors: A limited turing test for social-emotional intelligence. In Alexei V. Samsonovich and Valentin V. Klimov, editors, *Biologically Inspired Cognitive Architectures (BICA) for Young Scientists*, pages 34–40, Cham, 2018. Springer International Publishing.

[7] Matthew Crosby. Building thinking machines by solving animal cognition tasks. *Minds and Machines*, 30(4):589–615, 2020.

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 07 2011.

[9] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015.

[10] Ellen Goeleven, R. De Raedt, L. Leyman, and B. Verschuere. The karolinska directed emotional faces: A validation study. *Cognition and Emotion*, 22:1094 – 1118, 2008.

[11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[13] Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[15] Forrest Iandola, Song Han, Matthew Moskewicz, Khalid Ashraf, William Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size. 02 2016.

[16] J. Joo, W. Li, F. F. Steen, and S. Zhu. Visual persuasion: Inferring communicative intents of images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 216–223, 2014.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] Zdzisaw Kowalczuk, Micha Czubenko, and Tomasz Merta. Interpretation and modeling of emotions in the management of autonomous robots using a control paradigm based on a scheduling variable. *Engineering Applications of Artificial Intelligence*, 91:103562, 2020.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[20] Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. Introducing the open affective standardized image set (oasis). *Behavior Research Methods*, 49(2):457–470, 2017.

[21] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[22] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[23] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM ’10, page 8392, New York, NY, USA, 2010. Association for Computing Machinery.

[24] John Mullennix, Jeremy Barber, and Trista Cory. An examination of the kuleshov effect using still photographs. *PLOS ONE*, 14(10):1–13, 10 2019.

[25] G. Olague, E. Clemente, D. E. Hernndez, A. Barrera, M. Chan-Ley, and S. Bakshi. Artificial visual cortex and random search for object categorization. *IEEE Access*, 7:54054–54072, 2019.

[26] K. Peng, T. Chen, A. Sadovnik, and A. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, 2015.

[27] Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 07 1992.

[28] V. I. Pudovkin. *Film Technique and Film Acting*. Bonanza Books, 1958.

[29] Christoph Redies, Maria Grebenkina, Mahdi Mohseni, Ali Kaduhm, and Christian Dobel. Global image properties predict ratings of affective pictures. *Frontiers in Psychology*, 11:953, 2020.

[30] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, page 14, 2015.

[32] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[33] A. M. Turing. I.Computing Machinery And Intelligence. *Mind*, LIX(236):433–460, 10 1950.

[34] Matthew D. Zeiler. Adadelta: An adaptive learning rate method. *ArXiv*, abs/1212.5701, 2012.

[35] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. Affective image content analysis: A comprehensive survey. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI’18, page 55345541. AAAI Press, 2018.