Modelling the Quality of Visual Creations in Iconoscope

Antonios Liapis, Daniele Gravina, Emil Kastbjerg, and Georgios N. Yannakakis

Institute of Digital Games, University of Malta
antonios.liapis@um.edu.mt, daniele.gravina@um.edu.mt,
kastbjerg@gmail.com, georgios.n.yannakakis@um.edu.mt

Abstract. This paper presents the current state of the online game Iconoscope and analyzes the data collected from almost 45 months of continuous operation. Iconoscope is a freeform creation game which aims to foster the creativity of its users through diagrammatic lateral thinking, as users are required to depict abstract concepts as icons which may be misinterpreted by other users as different abstract concepts. From users' responses collected from an online gallery of all icons drawn with Iconoscope, we collect a corpus of over 500 icons which contain annotations of visual appeal. Several machine learning algorithms are tested for their ability to predict the appeal of an icon from its visual appearance and other properties. Findings show the impact of the representation on the model's accuracy and highlight how such a predictive model of quality can be applied to evaluate new icons (human-authored or generated).

Keywords: Online Game · Human Creativity · Crowdsourcing · Deep Learning · Mixed-initiative Design · Computational Creators

1 Introduction

While creativity has been a source of awe since the ancient years, as an activity of the gods in us [15], modern-day scholars of creativity have established that creative skills can be taught [4]. Indeed, creativity is increasingly being considered as an explicit educational objective within formal education [2, 17]. From *LEGO* to *Minecraft* (Mojang, 2011), games have been fostering the creativity of their players in a multitude of ways (construction, exploration, storytelling). Focusing on the theoretical framework of *creative emotional reasoning* [18], re-framing (i.e. changing a routine for performing tasks or a pattern of associations between facts, emotions or actions) can be accomplished through an external stimulus that causes disruption. Re-framing leads to semantic, visual, and emotional *lateral thinking* [4], which respectively target a shift in conceptual structures, visual associations, and one's perception of the effect that a creative solution will have on others' emotional states. A game that is designed explicitly around these concepts would be more targeted in the type of creative processes it elicits. When the game is used in the classroom and with teachers' intervention, it can be a powerful tool in *teaching for creativity* [7]. When the game is played in the wild, however, it would be valuable if the system could predict the visual impact of certain creative outcomes to help provide more targeted stimuli for disruption.

This paper revisits the Iconoscope project which was designed and developed in 2015 [11] and expanded with a web interface in 2017 [10]. Conceived initially as part of the FP7 ICT project *C2Learn* (project No: 318480), the game was intended as one of several classroom activities for young learners, with the teacher acting as facilitator and moderator. In this first version of Iconoscope, players would share a mobile device to draw an icon so that their group members could not always guess which concept it represents. Towards the end of *C2Learn*, and in order to maintain a persistent platform for playing Iconoscope in the classroom or in the wild, the game was redesigned to be played individually on a website in which all user creations would be publicly displayed. The game was further expanded for the purposes of the Erasmus+ project *eCrisis* (Europe in Crisis) with adjustments to the drawing interface and the way concepts were presented. To further enhance its usefulness in the classroom, a Do-It-Yourself (DIY) version of Iconoscope was developed to allow teachers to customize the list of concepts. When played in a classroom, DIY Iconoscope allows educators to discuss and play with specific topics of *eCrisis* such as social inclusion and integration. However, this paper will not focus on DIY Iconsocope but instead will analyze the users' data from the 45 months that Iconoscope has been online.

Given the long-term use of Iconoscope within the classroom (as part of the *C2Learn* and *eCrisis* projects) and in the wild (as the game is available to all on a public website), a large dataset of icons has been collected. Users have also engaged with the public gallery of the Iconoscope website, rating how appealing they find each icon and guessing which concept it represents. Given the recent advances of machine learning, this rich dataset of diagrams and user feedback could be used to train computational models of users' visual styles. This paper takes the first step in this computational modelling task by building predictive models of the crowdsourced visual quality of icons, reaching accuracies as high as 80% when combining image data with metadata on the icon's colors and shapes.

2 Machine Learning for Visuals

Machine learning grants the ability to automatically detect patterns in raw data, such as images, text, or audio. While conventional machine learning methods were held back by the requirement that raw data should be transformed into a suitable representation via a handcrafted feature construction process [1], *deep learning* [5] can circumvent this by automatically learning these representations from raw data through a nonlinear composition of simple data transformations.

Convolutional neural networks (CNNs) [9] are deep learning models applied to 2-dimensional inputs such as images or time-series data. CNNs employ a sequence of two-dimensional trainable filters (convolutions), nonlinear activation functions and pooling operations on the raw input, resulting in a hierarchy of increasingly complex features. By design, CNNs are able to encode the spatial information of their inputs. Since their success in the 2012 ImageNet competition [8], CNNs have become the dominant approach for almost all visual detection and recognition tasks [20, 16, 13].

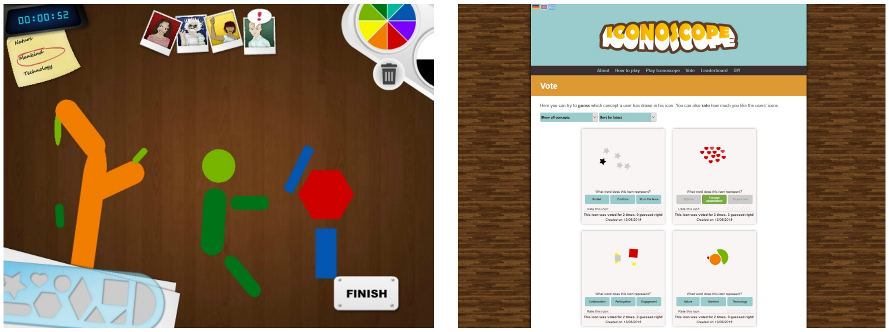
Deep learning has also been applied in multimodal learning tasks [22] which involve learning joint representation across multiple and (usually) heterogeneous information sources. Several studies have shown that combining multiple sources of information results in overall better performance, especially when the data size is limited or different tasks need to be learned simultaneously [14]. Literature in multimodal learning distinguishes between early and late fusion: early fusion aggregates information of different modalities via simple element-wise averaging, product and/or concatenation [14], while in late fusion high-level representations for each modality are computed separately and then fused via simple averaging or by stacking another learning model [19]. There is no consensus on which approach is better, and usually it depends on the task at hand.

This paper explores how to model the human perception of visual quality based on users' ratings. We test how different deep learning architectures, inputs and late fusion of different modalities (icons' images and metadata) affect the accuracy of a simple classification task (high rated vs. low rated icons).

3 The Game

Iconoscope is a creation game focusing on the visual depiction of semantic concepts in a creative fashion [12]. Creativity is fostered due to the constrained medium (as players must compose an icon from a small set of primitive shapes and colors) and due to the demand for ambiguity (as players must create icons that will be hard to guess). The game was initially designed for co-located play by a group of learners [11], but has been redesigned as an online game (*web*Iconoscope) which allows players to anonymously submit icons to a public database where all users can browse and provide feedback asynchronously [10].

Game Loop and Drawing Interface: In the online version of Iconoscope, players first select their language of choice (English, Greek and German translations of the game are available) and follow a tutorial. Players are then shown a list of different *concept triplets*, each in a different post-it note. Players select one of the three concepts in a concept triplet, and enter the drawing interface with the triplet and the chosen concept highlighted (top-left corner of Fig. 1a). In the drawing interface, players can add or remove shapes (among the available types on the bottom-left corner of Fig. 1a), change their color through the palette in the top-right corner of Fig. 1a, and move, rotate or resize them by dragging their relevant anchor points. At the same time, computational assistants can provide alternative icons to the user if the latter taps on one of their portraits (top of Fig. 1a). Assistants use computational intelligence methods to change the shapes and colors of the user's icon, and the user can choose to replace their creation with a computer-generated one or ignore it. Each of the four assistants has a different process for generating icons; more details can be found in [10]. When



Here you can try to **guess** which concept a user has drawn in his icon. You can also **rate** how much you like the users' icons.

What word does this icon represent?
Rate this icon:
This icon was voted for 2 times. 0 guessed right!
Created on 13/06/2019

What word does this icon represent?
Rate this icon:
This icon was voted for 3 times. 3 guessed right!
Created on 13/06/2019

What word does this icon represent?
Rate this icon:
This icon was voted for 2 times. 2 guessed right!
Created on 13/06/2019

What word does this icon represent?
Rate this icon:
This icon was voted for 2 times. 0 guessed right!
Created on 13/06/2019

(a) The Iconoscope drawing interface (b) The Iconoscope voting interface

Fig. 1: Interaction methods with the Iconoscope game and website

the player is happy with their icon, or after a maximum of 5 minutes, the icon is sent to the database where everyone can see it and provide feedback (see below).

Concept Triplets: Creativity in Iconoscope lies in the interpretation of a semantic concept through an abstract visual icon that can mislead others to pick different candidate concepts. This is achieved through *concept triplets*, i.e. groupings of three semantically linked concepts, each consisting of one or more words. The three concepts can be linked by lexical similarity (e.g. "Lead, Govern, Dominate"), common social issues (e.g. "Sexual orientation, Gender, Human rights") or overarching theme (e.g. online identities in "Avatar, Communication, Expression").

Public User Feedback Interface: All icons are stored in a database and shown on the Iconoscope website[1]. As with the entire website, the "voting" page (see Fig. 1b) is accessible to any visitor, who can offer feedback anonymously on any of the icons in the database. Under each icon, the three concepts of the triplet selected by the icon's author are shown. The user can attempt to guess which of these three concepts is depicted, and can also see how many other users have guessed correctly. Once the user chooses one of these three buttons, they receive feedback on whether they were right or wrong and can not guess again for this icon. The website allows each IP to make one guess per icon. Additionally, users can "rate" how much they like each icon, on a scale between 1 and 5 stars. Similar to guessing the concept, one rating per icon is allowed from each IP.

Since icons created via Iconoscope are intended to be ambiguous and difficult to guess correctly, the website includes a leaderboard which shows the top 10 icons with the highest ambiguity score. This ambiguity score is calculated based on a balance between correct and incorrect guesses of users, while also rewarding icons which have received more guesses in total; see [10] for more details.

[1] `http://iconoscope.institutedigitalgames.com/vote.php`

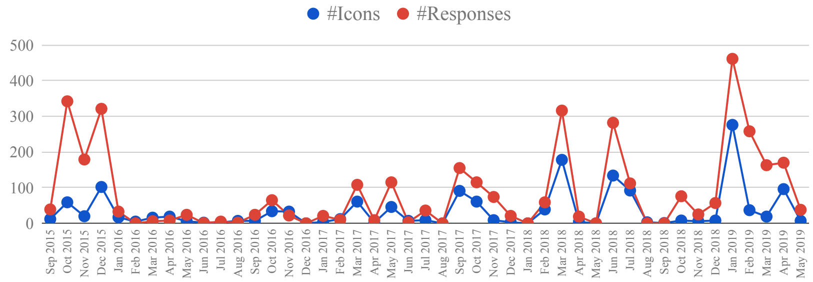


Fig. 2: Monthly interactions with the drawing interface and the voting interface.

4 Data Collection

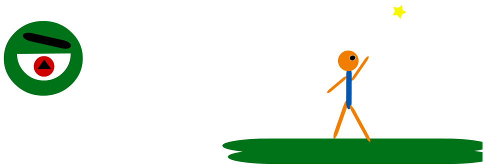
The online version of Iconoscope was launched in September of 2015. This Section analyzes the data collected from Iconoscope since its launch until 23 May 2019, which is 45 months (almost 4 years) of continuous operation.

Icons created and user feedback: After curation by educators and administrators for offensive content and cleanup of corrupt data, a total of 1555 icons were collected during the 45 months examined. These icons received 3774 user responses through the website's gallery: these responses include guesses, ratings, or both. Fig. 2 shows the distribution of these interactions over time.

Favored concepts: In terms of the concept triplets and the selected concepts among them, it is not surprising that some of the concepts were favored more than others for creating icons. The most popular triplet was "Nature, Mankind, Technology" (164 icons), likely because all three concepts seem intuitively straightforward to draw. In contrast, the least popular triplets are "Proactive, Reactive, Inactive" (24 icons) and "Tolerance, Acceptance, Solidarity" (30 icons) which consist of much more abstract concepts. As expected, the most commonly depicted concept was "Nature" (in 102 icons) followed by "Push" (76 icons), "Play" (70 icons), and "Team" (65 icons).

Icon properties: The icons collected during the 45 months of Iconoscope contained a total of 7912 shapes. While this amounts to 6.1 shapes per icon, icons most often had one shape (14%), two shapes (11%), three shapes (12%) or four shapes (10%). However, 27 icons included 20 or more shapes and the highest number of shapes in one icon was 119. In terms of types of shapes favored, circles were most common (22% of all shapes) followed by elongated rounded rectangles (17%) and squares (14%). From observations of created icons, rounded rectangles were often used as lines (as Iconoscope does not include lines).

In terms of the icons' colors, it is not surprising that 30% of icons only had one color (as 14% of icons had one shape in any case). Most other icons had either two colors (23% of all icons), three colors (18%) or four colors (12%),



(a) Icon for the triplet "Danger", "Safety", "Protecting the young"

(b) Icon for the triplet "Nature", "Mankind", "Technology"

(c) Icon for the triplet "Protest", "Conform", "Sit on the fence"

Fig. 3: Some of the icons rated with the maximum score (5); these icons had the highest number of ratings. In total, 142 icons received an average rating of 5.

although there were also several icons with 10 or 11 colors (9 icons and 4 icons, respectively). While all possible colors in icons are 11, only 10 are options on the interface (eight hues, black, and white). New shapes start as gray, but players do not have an option to recolor shapes to gray. Gray shapes were surprisingly common (32% of all shapes); it is obvious that users often chose not to recolor their shapes. Other popular colors were red (14% of all shapes) and yellow (10%).

Playtime: In terms of the time users spent drawing each icon (i.e. *playtime*), the average playtime was 133 seconds. Users seem able to finish icons fairly quickly (28% of icons were created in under one minute). This is not surprising, considering that many icons had one or two shapes. Indeed, there is a strong positive correlation between the number of shapes and playtime (Pearson correlation coefficient $\rho = 0.47$). On the other hand, 12% of icons were submitted automatically by the system when the time ran out (5 minutes); this indicates that some users could not identify how to submit icons manually in the interface.

Use of assistants: It is worthwhile to investigate how users interacted with the four included computational assistants who serve as aides to players' creativity. In total, users selected assistants to receive their suggestions 2575 times; out of those, users applied the suggestions to replace their own icon 499 times. The assumption is therefore that only in 19% of instances were the computational assistants' creations considered helpful—or, to be precise, better than the user's own sketch. A confounding factor, however, is the fact that assistants animated and showed a dialog balloon at random intervals, which may have urged users to select them even when they did not want to change their design.

Public feedback: As noted above, a total of 3774 responses were collected from the public gallery of the Iconoscope website. Out of those, 835 responses included a rating of the icon in terms of appeal. Such ratings were only offered on 521 of the 1555 icons; it can be assumed that only some of the icons captured the attention of the audience enough to receive ratings (even if that rating was bad). Figure 3 shows a sample of the highest rated icons for different concept

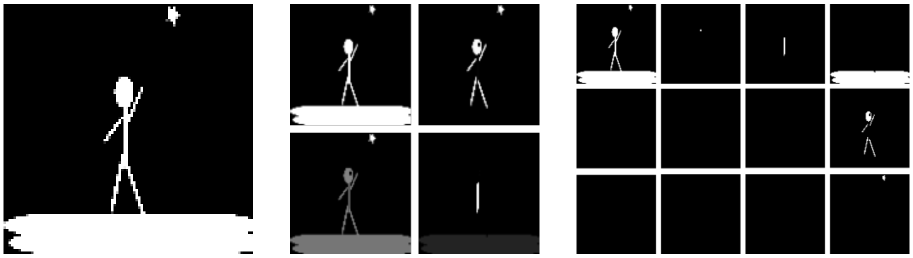

(a) Alpha channel (b) ARGB channels (c) 12 binary channels

Fig. 4: Example pre-processing of the image inputs for the image of Fig. 3b.

triplets. Moreover, 3710 out of 3774 responses included attempts at guessing the depicted concept. Most icons were annotated in this fashion, as 1370 of the 1555 icons (88%) received at least one attempt at guessing the concept. Most icons received one or two guesses, with only 34 icons receiving more than 10 guesses.

5 Modelling Visual Quality

As noted in Section 1, the corpus of icons and user feedback can be used to train computational models via machine learning. As a first exploration in this vein, this paper focuses on training models that can predict an icon's visual appeal, using the crowd's ratings of appeal as the ground truth. The application of such a model could be to predict the visual quality of icons that have not received ratings (less than a third of icons have received any rating on appeal). More ambitiously, such a model could be used by the computational assistants which attempt to generate alternatives to the user's icon; instead of or in tandem with the assistants' current objectives, assistants can attempt to improve the predicted visual quality of their generated icons. The next sections discuss how the data is prepared and the results of different machine learning experiments.

Preprocessing: For the task of predicting the rating of icons, the dataset consists of the 521 icons that received at least one rating from users, coupled with the average value of users' ratings for each icon. In order to simplify the task of predicting visual quality, we treat it as a *binary classification* task between "high" and "low" rated icons. To assess where the split between high and low rated icons should be, the average μ of all icons' ratings is calculated ($\mu = 3.15$). In order to avoid ambiguous annotations, the icons with an average rating within 5% of μ, i.e. $[2.99, 3.31]$ are ignored; 84 icons in total are ignored in this fashion. Icons with an average rating below 2.99 are treated as low rated, and icons with average rating over 3.31 are treated as high rated. This split yields 247 high and 190 low rated icons. To validate our machine learning findings, we apply 10-fold cross-validation and apply oversampling to the least common class in the

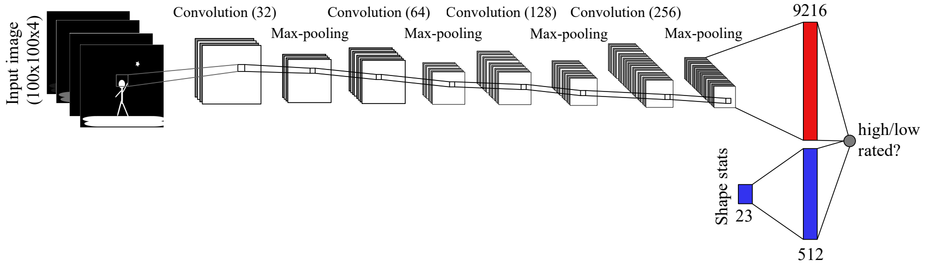


Fig. 5: Late fusion architecture for image channels (here: ARGB) and shape stats.

training and the test set individually; this results in 445 training samples and 51 testing samples on average. The baseline accuracy (random choice) is 50%.

This paper explores several ways of processing the icon to be used as input for the machine learning task. The 2D image of the icon is the most straightforward input. Each image is cropped to the icon's bounding box (removing unnecessary empty space) and scaled to 100 by 100 pixels. After this, the scaled-down image is treated here in four different ways (see Fig. 4): as a binary *alpha* channel (transparent versus non-transparent), as RGB channels, as ARGB channels (including transparency information) and using a custom 12-binary channel format based on the 11 possible shapes' colors in Iconoscope and the alpha binary channel. In addition to the image, we test the *shape statistics (stats)* of the icon as another modality of input. The shape stats make up a vector of 23 real numbers listing the number of primitives for each possible shape (10), the total number of different shapes (1), the number of primitives for each possible color (11), and the total number of different colors (1).

Results: A number of machine learning structures and input modalities were tested: Table 1 shows the results of these experiments. All networks end in one output node which predicts high (1) or low (0) rated icons. When using shape stats as the only output, an artificial neural network (ANN) is used with a single fully-connected hidden layer of 512 nodes. When using only image inputs, the output of the convolutional network is flattened into a vector and connected to the single output. When combining images with shape stats, the flattened vector from the CNN is concatenated with the 512 nodes which process the shape stats (late multimodal fusion, see Fig. 5). Based on extensive parameter tuning, the CNN we use has four layers of convolution (of size 5×5 with zero padding), with 32, 64, 128 and 256 filters, each followed by a max-pooling layer; this results in a flat vector of 9216 features. Finally, we tested a pre-trained VGG19 [20] which is a very deep architecture trained on the vast ImageNet image corpus; the VGG19 produces a flat vector of 4608 features which is concatenated with the 512 nodes from shape stats or fed directly (image-only) to the output. The VGG19 model accepts RGB images only, and training is only applied to the final layer's weights. All nodes use an ELU activation function [3] and the output of each hidden layer is normalized via batch normalization [6]. All models were trained for 20 epochs,

input	image	image & shape stats
Shapes only (NN)	68.54 (6.26)	
Alpha (CNN)	64.27 (5.31)	69.40 (5.94)
RGB (CNN)	64.71 (5.55)	67.53 (6.29)
ARGB (CNN)	64.18 (5.04)	67.88 (5.97)
12 channels (CNN)	64.45 (4.79)	68.95 (5.36)
RGB (VGG19)	68.12 (5.35)	68.33 (6.21)

Table 1: Test accuracies (%) for different networks and inputs, averaged from 10-fold cross-validation. Standard deviation across folds is shown in parentheses.

while to avoid overfitting we save the best model obtained during the training process based on validation accuracy. Since the training data is sparse, we use dropout after each hidden layer [21] and reported results in each case are the best across three different dropout values tested (0.1, 0.3, 0.5).

Based on Table 1, we notice that including shape stats generally increased performance. The exception is VGG19, which had comparable accuracy with or without shapes (best performing fold is with images alone: 77%); since VGG19 is trained on millions of real-world images, it is not surprising that it is strong in visual pattern detection. Surprisingly, using shapes alone as input achieves comparable accuracies to CNN models (best fold: 81%). Overall, all CNNs were well performing when combining image data with shape stats as input, with the alpha channel component achieving the highest accuracy on average (69%) while the ARGB component reached the highest accuracy at the best fold (83%).

6 Conclusion

Results of Section 5 show that average accuracy when predicting visual quality is not very high, but metadata regarding shapes' types and colors can help in that regard. Current models of visual quality (e.g. the best fold) can be used to rank all icons in the corpus, even if these have not received any ratings from users, or even to predict visual quality for new icons as they are created. Predicted visual quality can also be used as a constraint for the computational creators' search processes, ensuring that their suggestions are at least predicted to be high rated. Future work should explore other deep learning models using the icons (and shape stats) as input to predict metrics such as the ambiguity score (based on users' guesses), playtime, or—more ambitiously—the concept being depicted.

References

1. Ballard, D.H., Hinton, G.E., Sejnowski, T.J.: Parallel visual computation. Nature **306**(5938), 21 (1983)
2. Cachia, R., Ferrari, A., Kearney, C., Punie, Y., Van, W., Berghe, D., Wastiau, P.: Creativity in schools in Europe: A survey of teachers. http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=2940 (2009), accessed: November 2016

3. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (ELUs). arXiv preprint arXiv:1511.07289 (2015)
4. De Bono, E.: Lateral thinking: Creativity step by step. Harper Collins (2010)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
6. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
7. Jeffrey, B., Craft, A.: Teaching creatively and teaching for creativity: distinctions and relationships. Educational Studies **30**(1), 77–87 (2004)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (2012)
9. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10), 1995 (1995)
10. Liapis, A.: Mixed-initiative creative drawing with webiconoscope. In: Proc. of the Intl. Conf. on Computational Intelligence in Music, Sound, Art and Design. (EvoMusArt), vol. 10198, LNCS. Springer (2017)
11. Liapis, A., Hoover, A.K., Yannakakis, G.N., Alexopoulos, C., Dimaraki, E.V.: Motivating visual interpretations in iconoscope: Designing a game for fostering creativity. In: Proceedings of the Conf. on the Foundations of Digital Games (2015)
12. Liapis, A., Yannakakis, G.N., Alexopoulos, C., Lopes, P.: Can computers foster human users' creativity? theory and praxis of mixed-initiative co-creativity. Digital Culture & Education (DCE) **8**(2), 136–152 (2016)
13. Makantasis, K., Doulamis, A., Doulamis, N., Psychas, K.: Deep learning based human behavior recognition in industrial workflows. In: Proc. of Intl. Conf. on Image Processing. pp. 1609–1613. IEEE (2016)
14. Park, E., Han, X., Berg, T.L., Berg, A.C.: Combining multiple sources of knowledge in deep cnns for action recognition. In: Proc. of the Winter Conf. on Applications of Computer Vision (WACV). pp. 1–8. IEEE (2016)
15. Plato: Ion. In: Plato: The Collected Dialogues. Princeton University Press (1961)
16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. of the IEEE Conf on computer vision and pattern recognition. pp. 779–788 (2016)
17. Sawyer, K.: Educating for innovation. Thinking Skills and Creativity **1** (2006)
18. Scaltsas, T., Alexopoulos, C.: Creating creativity through emotive thinking. In: Proceedings of the world congress of philosophy (2013)
19. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems (2014)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
22. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in neural information processing systems (2012)