From Visual Forms to Metaphors – Targeting Cultural Competence in Image Analysis

Lars Oestreicher [1] and Jan von Bonsdorff [2]

[1] *Department of Information Technology, Uppsala University, Box 337, 751 05 Uppsala, Sweden*
[2] *Department of Art History, Uppsala University, Box 630, 751 26 Uppsala, Sweden.*

Abstract

Image analysis has taken a large step forward with the development within machine learning. Today, recognizing images as well as constituent parts of images (faces, objects, etc.) is a relatively common task within machine learning. However, there is still a big difference between recognizing the content of a picture and understanding the meaning of the image. In the current project we have chosen an interdisciplinary approach to this problem, including art history, machine learning and computational linguistics. Current approaches pay large attention to details of the image when trying to describe what is in the picture, resulting, e.g., in that smiling faces will support the interpretation of the image as "positive" or "happy", even if the picture itself is a scary scene. Other problematic issues are irony and other polyvalent messages with a large amount of ambiguity that enables for example humorous interpretations of a picture. As a starting point, we have chosen to identify visual *agency*, i.e., how and why pictures, when regarded as acting agents, effectively may catch the attention of the viewer.

Our objective for this first phase of the project is to investigate multi-modal models' capacity for recognizing such high-level image content as, for example, *context*, *agency*, *visual narration*, and *metaphors*. Ultimately, the goal is to improve cultural competence and visual literacy of neural networks through art-historical and humanities expertise. In the paper we will describe our current approach, the general ideas behind it, and the methods that will be used.

Keywords

Multi-modal machine learning, high-level image content, visual metaphors, cultural competence, pictorial conventions

1. Introduction

Gottfried Boehm tells us that the image includes a duality; it shows away from itself, while still maintaining a materiality: "Bilder sind spannungsgeladene, real-irreale Körper" (Images are bodies fraught with tension, simultaneously material and immaterial) [1]. In this way, it is more or less obvious that a picture is in some sense "larger" than the bare sum of its parts. Let us call this the Jack-in-the-box quality. When you open the "box" of the image, you tend to get more than you ask for.

This is our conception of the image: Any well-conceived image is an active interpellative instant, speaking, violently interrupting, yelling, and tugging at the observer's sleeve. For us the image or the work of art can work as a kind of golem, a being with restricted life-like properties. Still, if this golem wants to communicate, it has to use some kind of sign system known to others. It is this systematized will-to-communication that we want to single out and convey to the neural networks.

Today, we have a new player in the field of image interpretation, namely systems based on artificial intelligence, primarily in the shape of machine learning within the area of image analysis. The systems that are being developed can give more and more precise descriptions of the content in the pictures, in terms of what objects the image contains. It is even possible to see physical relations (even in depth,

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
Email: LarsOe@it.uu.se (A 1); Jan.Von.Bonsdorff@konstvet.uu.se (A 2)
ORCID: 0000-0003-2560-425X (A 1); 0000-0002-9482-9690 (A 2)

such as distance) or even shallow relationships [2] between the objects in the image. By shallow, we mean such that can be described by simple expressions, such as “Man wears glasses” or “Woman holds spanner”. Even so, the more higher-level message of the image is still not addressed by these systems. These shallow interpretations still reside on a relatively simple level, mostly providing the superficial interpretation of the objects as mentioned above. In this way, most machine learning applications would for example have serious problems making sense of some of the more complex pictures by Albrecht Dürer (see Figure 1). Identifying the objects in the picture is quite simple (as long as the image recognition software has been trained on the type of objects that are involved in the picture).

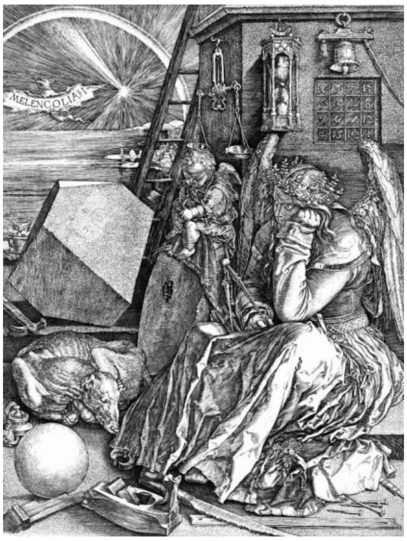

Figure 1. Albrecht Dürer, Melencolia I (Engraving 1514).

For example, it would be relatively easy for an artificial intelligence to identify discrete, well-known elements in the engraving: the two angels (one large, one small), the dog, the scientific instruments, the ladder, the rainbow. Projects like for example Saint George on a Bike (Barcelona Supercomputing Center) have shown a working model of making discrete iconographic motifs in medieval fine art accessible to machine learning and artificial intelligence [3]. What is more, the open Web platform iArt (https://labs.tib.eu/iart), specializes in differentiated fine art searches, using a complex modular system architecture [4]. The iArt search engine masters iconographically based classification principles that, e.g., examine objects for biblical motifs or general genre themes.

But back to Dürer’s engraving, where uncertainties appear at all levels. What about the planet in the background – or is it a comet? What kind of space does the angel reside in? Is it a terrace high above sea level? Do the architectural elements imply an open or a closed room? What kind of geometrical form is the weird polyhedron? The AI could perhaps work this out, if equipped with a specialized data base. The same applies for the rich and well-documented iconographic tradition of personifications and allegories of Melancholy and Geometry. But what about making a relevant synthesis of the vast corpus of literary and scientific knowledge connected with the concept of Melancholy and more specifically with the engraving, trying to reach the level “what is it all about”? Typical confident statements, at times summarizing years of art-historical studies, such as “This engraving is an allegory of the creative mind in repose” or “This is a mind-map of Dürer's scientific thinking” or “The engraver is fooling the audience with nonsensical content” are certainly beyond the current capacity of the machine. We do not strive to reach any confident closures, but rather a tool that can make worthwhile suggestions, see unexpected connections and follow mental leaps.

2. Theoretical Background

The project depends heavily on both the latest developments within artificial intelligence and current theoretical trends within the field of art history. The intersection between these areas has attracted much attention lately, where the AIA is discussed both as producing artist and as analytic perceiver. The perception of an image by a machine learning system still acts on a fairly low level, and there is a lack of research on the deeper levels of the understanding/interpretation of art in this area, which is the focus of the project presented in this paper.

2.1. Art History and Meaning

Most images do not only portray a certain scene, but rather try to tell a (visual) story, and our approach does not start on the formal level of object detection, but commences at the instance when images start to do more than just show things, i.e., when they engage the observer and start to tell stories. The visual stories are difficult to describe in words, since the images are so visual in their character that linguistic expressions do not seem capable of capturing an adequate meaning of all the aspects of an image. Many aspects of an image are also dependent on the background of the observer, his or her background, cultural context etc. For example, the interpretation of a painting with a religious scene benefits from an understanding of the religious belief that lies beneath the motif of the painting [2]. Different details of a painting will trigger different associative patterns depending on whether the image is viewed from the cultural perspective or from a more direct interpretation of the scene. Of course, the respective details of the picture can be described with a similar understanding of the elements. However, the interpretation of the visual story will most likely differ significantly.

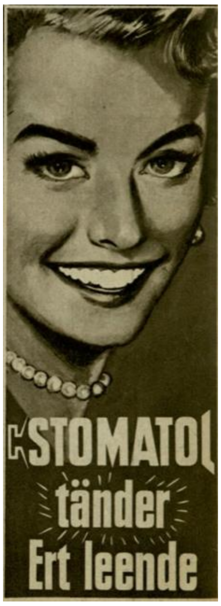

Figure 2. A first sample advertisement from the material collected in the pilot study (from 1956).

If the interpretation framework is not known, the visual story or narrative will often not be recognised, or only recognised with large difficulty [5]. Of course, all possible contexts of an image cannot be known. The amount of unknown and unreadable contexts seems to grow with distance in time. Cave paintings from different periods as well as Neolithic rock art can seem impenetrable except in the most general meaning. Much of the beliefs and incentives of these long-lost cultures are known only through the visual recordings with no possibility of comparative corroboration through other historical sources.

Interestingly enough, the difficulty of the close reading of a culture also pertains – to a degree – to our chosen visual sources, the visual ephemera of advertisements from the 20th century. Some content may be readily identified, like the smiling lady in the tooth paste ad from 1956 (See Figure 2). The visually stated metaphor (to be precise, a simile) reads: “The lady's teeth are like pearls”. In this connection, it seems fruitful to turn to conceptual metaphorics as formulated by, among many others, Lakoff and Johnson [6]. Visual metaphors are wonderfully suggestive and effective in engaging the viewer’s interest and starting up narratives of the kind we are looking for; the viewer has to make mental leaps, and the effort of the resulting combinations seems to bring the audience immediate reward in terms of an expected story. This suggests a very swift accumulation of meaning through nearly instantaneous interpretation. This is the “Jack-in-the-box” quality we mentioned earlier.

The attribute is a feature typical of visual metaphors. The attribute can be an aspect of the interpreted thing itself: “white” teeth can be made “pearly and lustrous”. But usually, a metaphor (verbal or visual) contains two domains, that are not the same but contain overlapping features, like Shakespeare’s famous metaphor of the “eye of heaven”, that is, the sun.

In this humble, but quite suggestive toothpaste advertisement, “teeth” are given positive attributes; they are “shiny” as the light (the Swedish word “tänder” can mean both “light up” and “teeth”). Providing the woman with well-known attributes fixes the firm staging of a possible story-to-be-told. The pointers to strong connotations work as a kind of semantic reinforcement or amplification [5]. When the object is clearly depicted with familiar attributes, it aids the identification, it securely establishes the objecthood of the thing in the physical world, and, in this case, it is meant to impress. The agent (here the young lady) is endowed with a stronger presence when bestowed with amplified characteristics, here “lustrous, shiny teeth”.

This is as far as we can go with the general content of the image. What cannot be known without other sources, is that the attributes of “light” and “gleam” also alludes to a real-world fact: The tooth paste brand “Stomatol” was the text of the first electric sign advertisement in Stockholm in 1909, a circumstance still well-known to the readers of Swedish magazines at the time of the advertisement from 1956, but not so familiar today.

Thus, some content is more “general”, some “special”. In our annotation practice we may not be able to grasp some of the more specialized meanings. In these cases, the general meanings will have to do. But – we will come a long way with specialized art historical background knowledge and terminology, so we feel well equipped to tackle the aspect of content loss through time.

2.2. Artificial Intelligence for Meaning Detection

In the area of artificial intelligence, the project will utilize a combination of methods from image analysis and natural language processing. We will in the following look at each area and then discuss the possibilities in a combined method. Starting from a set of annotated images, the main idea is that the system will be able to perform an analysis of the annotations as a means to categorize the images in themes or topic areas. The topic areas can then be anchored onto the themes in terms of, for example, a more pragmatic interpretation (in the end even considering metaphorical meanings in the images).

2.2.1. Computational Image Analysis

The use of computers within the visual Arts field has been researched almost as long as computers have had the capacity to calculate large arrays (of pixel values). The programs at that time has mostly been as a supportive tool for artists, in the shape of, for example, Adobe Photoshop, and the more artistically oriented Corel Painter application. The issue of image generation and manipulation has also been a popular field of development in the early 1980s, mostly with miniature programs such as “Pico” [7]. Computers were from a very early time considered as very potent actors as producers in the Visual Arts (see, e.g. the volume by Spalter [8]). However, the capacity for automated analysis of images was still lacking, both in terms of hardware and software resources.

With increasing computational capacity, the area of computational image analysis gained a large interest, especially within medical applications, for example as a tool for early diagnosis of illnesses, through the early identification of cell anomalies, such as cancer cells. Initially this analysis was based on traditional software solutions, using large statistical packages.

Image analysis has grown into an expansive area of research already before the introduction of deep learning. The application area has grown and the research within general image analysis has taken a large step forwards in the last decades, much due to the progressive development within *machine learning and deep learning,* currently reaching image recognition and object detection precision of more than 90 percent in the average case. As an example, we will look at the simplest principles for image analysis, where the analysis is based on the recognition of gradually more and more complex features that are found in the image. The algorithms used in this type of systems depend on finding features in the pictures on higher and higher levels. The analysis methods applied here are basically of a statistical kind, but in a less guided form than traditional statistical analysis methods.

While training a machine learning model, the analysis is directed towards finding possible patterns at different levels of abstraction in the training data (the pictures), which in itself are reduced to matrices of numbers, representing images, sentences, data structures or any other kind of data. This means that the training and use of a ML-system, in one aspect, is insensitive to the type of data, but only tries to find recurring data patterns in the numbers provided as input.

In the lower levels the simplest features will be recognized as, e.g., lines or points which are then clustered and recognized at higher and higher levels of abstraction from the basic pixel level. In a mid-level feature detector, we find some shapes that can be seen in the image as belonging to parts of the motif, such as wheels (circular shapes) and sweeping arches, which are then even more recognizable in the high-level feature map.

The classifiers then detect the features in the images, and will combine features into clusters that signify object types in an image. Through this principle the networks, when trained, can detect different kinds of objects in the images, and through these also draw conclusions of image themes, such as whether the picture shows the already mentioned big city or a mountain village, a harbour or a supermarket parking, for example.

Through these methods of feature analysis, it is also possible to train networks for an analysis of anomalies, such as, the method used to detect cancer cells in tissue samples, or for the detection of retinal changes in case of certain diseases. There are a large number of application areas where this type of analysis has brought huge possibilities to the field. There are also constant improvements to these application areas from public competitions, such as those managed on the competition oriented Kaggle website (https://www.kaggle.com). Today the simpler methods for image analysis with Deep Learning can be applied successfully even as hobby programming, and what were previously research problems are today often not even suitable for theses on the lower levels of education. However, it is also a fact that more complex image analysis applications still require a large amount of work, both in programming, as well as runtime processing time.

The application of machine learning to art in general, and more specifically to imagery is also an expansive field, with quite a few examples of interesting applications, such as the detection of art style, and the re-application of art styles to pictures and even to photographs, for example in the shape of numerous applications for photo manipulation being available in mobile devices. These techniques are now so wide-spread that this has become a large ethical issue with, for example, fake pictures of celebrities being spread [9]. With this plethora of applications for machine learning in combination with images, we are tempted to assume that most of the problems have been approached in one way or another by machine learning applications.

When we look at the approach as outlined above, most of the image analysis and manipulation applications are still based primarily on a massive statistical analysis of features. But, image analysis and interpretation from the perspective of art is much more complex than just detecting denotative features or the general drawing style [10]. In the images there are other subtler communicative aspects that are essential to the interpretation of the pictures, but which are difficult to catch with the current image analysis methods. These less obvious nuances may even sometimes not be covered by a sampling of the objects in the picture, or by the manner of painting, but depend on more subtle aspects of the picture. In fact, some of the factual clues to an interpretation of a picture may not even be possible to specify in words or expressions. However, human observers are often able to classify images according to their message from this perspective.

2.2.2. Natural Language Processing for Images

Also, within the research in computational linguistics, there has been a large development with the introduction of machine learning. Previous systems were mostly based on a use of semantic or conceptual network representations of knowledge, where the texts were translated into large, interconnected networks of nodes (see for example [11], [12]). More grammatically oriented attempts were also used, for example, based on case grammar descriptions [13]. These approaches have more or less disappeared, in favour of more statistical method. Many early language translation systems used statistical methods as a base. The precision and capacity of these early systems were low. With increasing hard- and software capacities, the precision improved greatly over time.

A major change came with the introduction of neural networks, and the field has expanded into many different directions, from language translation to more specific systems, used for summarizing, question answering and sentiment analysis (for example, of movie reviews). Many of these systems make use of pretrained language models, which can be fine-tuned to be used for new tasks (also referred to as transfer learning). A prominent, recent example of such systems is the Bert system [14], which aims at building a model for language understanding in different domains. A later development resulted in the Visual-linguistic Bert (VL-Bert), which combines methods for image analysis with the Bert system [15]. VL-Bert is supposed to facilitate visual common-sense reasoning as well as answering questions about images. This is one step towards the higher-level analysis that is the goal of our project.

Another major change is the event of transformer-based, multimodal neural networks building on massive amounts of raw data. In January 2021, OpenAI introduced CLIP (*Contrastive Language-Image Pre-Training*) [16]. CLIP is a neural network which efficiently learns visual concepts from natural language supervision. It learns directly from raw text about images (without resorting to manually labelled data). CLIP relies on 400 million image-text pairs from the internet, that is, images with captions. CLIP models can be applied to nearly any visual classification task without needing additional training examples (like ImageNet would). And furthermore, CLIP allows researchers to design their own classifiers and removes the need for task-specific training data.

The significance of the approach used in CLIP lies in that, in contrast to most other systems, it will not only tell which class a certain object in the image belongs to, but it can provide an adequate result on a natural language search prompt of what the image depicts. In this way it will add some (still superficial) contextual knowledge to the picture, as one step in the direction we would like to work.

We will not go into details about the technical aspects of the CLIP system here, but there are some features that are worth noting. The system has for example been trained on text/image pairs, such as could be gained from social media, such as Instagram, where people are publishing photos *together* with their own taglines, which often describes the motif in more detail, than just “a cat”. The system also does not “kill” the connotations through the creation of distinct classes (which can be translated into non-meaningful, but computable entities, such as numbers or token strings).

The CLIP system is built in such a way that it will provide a description which is close to natural language, or, provide a picture that shows what the description entails. It is a system that can work in two directions, either providing textual descriptions from images, and providing images that are found from textual descriptions. According to the company openAI, the system contains of two types of encoders (one for images and one for texts) in combination with “zero-shot transfer, natural language supervision, and multimodal learning” (https://openai.com/blog/clip/).

It is in this context important to note that neither CLIP nor the natural language application GPT-3 (produced by the same company) can be said to really *understand* the expressions inherent in the images or the texts that it produces or analyses. The output of this type of system is essentially a statistical prediction, based on the large number of calculations made over huge amounts of data samples. In short, the system analyses a very large amounts of pictures together with descriptions of the pictures made by humans. From this input data it generates a prediction of what would be the most likely description given by a human to the new images. However, it cannot be said to analyze and interpret neither the images, nor the texts *in the human sense*. The situation is almost identical to the dilemma described by Searle, in his “Chinese room” example [17], where he, in short, states that if an entity (human or machine) learns and uses a large number of rules for the handling of language, it can produce perfect sentences and even answers to questions without understanding a single word in that language. Thus, the understanding of the visual stories in the pictures, as described previously in this paper, are still way beyond reach with the current technology.

With this background, it is of course also relevant to question whether it will be possible for a machine learning agent to perceive not only the primary objects in the image, but also create a representation of the visual story being told? How much of the interpretation of an image and of its visual story relies on an actual understanding of the context and the symbolic meaning? This is where our project takes its starting point.

3. Method and Material

What kind of methods and models do we need when annotating the visual material? We have already mentioned our choice of starting the description with the notion of agency and the beginning of a story ("narrative kernel") often found in visual metaphors. A feature of visual metaphors is the one we mentioned already as "the jack-in-the-box quality": Images – or rather, their interpretations – tend to start from a latent quality, a mere potential, to emerge as qualified meaning-carriers.

Let us present one example on the drift into a level of denser signification from our experiments on exemplary image annotations: In many advertisements, a wide-angle, one-point perspective view may often be used as a trope for "modernity and progress" (See Figure 3). This ad for a German car brand ("Build a Bridge over Bad Roads!") shows a bumpy sand road and a conceptual grey bridge in curvilinear orthogonals, converging to a vanishing point on the left hand. Speed lines also converge at the vanishing point, hinting that the car is quickly closing in from a great distance. The car is depicted in a worm's-eye view. The sketch of the axle track is inspired by technical drafting and adds to the techy impression. This kind of scenery using steep linear perspective is a kind of trope for societal advances.

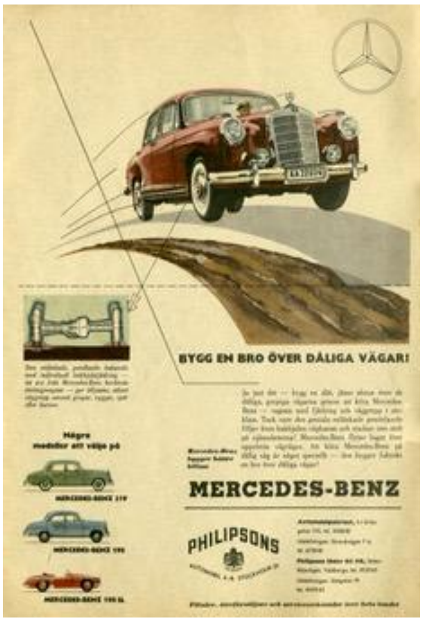

Figure 3. A second advertisement (published 1958) from the material collected in the pilot study.

In Figure 4 two other examples are shown: In the 50's, the Shell company presented a series of ads about different collaborations with other companies, here a cable and wire company. The ads are high-quality, with a weird fusion between Surrealism and the depiction of technology as a road into the future. The undated tie ad (Raxon Fabrics) is not as skilfully devised as the cable ad, but nonetheless interesting in its use of spatial cues. The forced perspective has no apparent role in suggesting space for the ties hanging in a conceptual void. The orthogonals with the pole as a kind of vanishing point constitute more like a sign on its own, a sign of progressive times and the future. Through interpretation of images, meaning seems to slide from the uncategorized formal feature into a region where storytelling can begin. Thus, we need not a rigid, hierarchical model as a basis for our methodology, but a model that allows for movement, opening, and emergence between levels of meaning. Until now, we have gathered a smaller corpus of randomly selected issues from the richly illustrated monthly and weekly magazines of for example *Veckojournalen* and *Bonniers Månadstidning* from the 1920s to the 1950s, a dataset of approximately 12,600 discrete images.

4. A Combined Approach

Currently, AI-based search engines are less knowledgeable about image content and context than what is desirable for meaningful performance ranking. Therefore, in the first stage of our pilot project, we want to apply multi-modal machine learning models that can connect text and images, like CLIP, in an image database. The data consists of visual ephemera like advertisements and photo journalism from Swedish weekly and monthly journals from the 1920's to the 1950's, as well as assorted images of fine art. In a later stage, we want to test different kinds of specialized image annotations in natural language. Thus, our **objective** is to investigate the multi-modality models' capacity for recognizing such high-level image content as for example context, agency, visual narration, and metaphors. We are also interested in how a space built from such pair-wise distance would look and its properties in relation to qualitative theory.

The **first problem** we want to tackle is finding the most effective method of handling a large number of annotated and unannotated images. Many current approaches contain major obstacles: typically, visual datasets, often based on crowdsourcing, are labour intensive and costly to build. Furthermore, the visual concepts and classifiers the datasets are able to learn are narrow and difficult to supplement. Connected to this first problem is also the extraction of our image-and-text datasets.

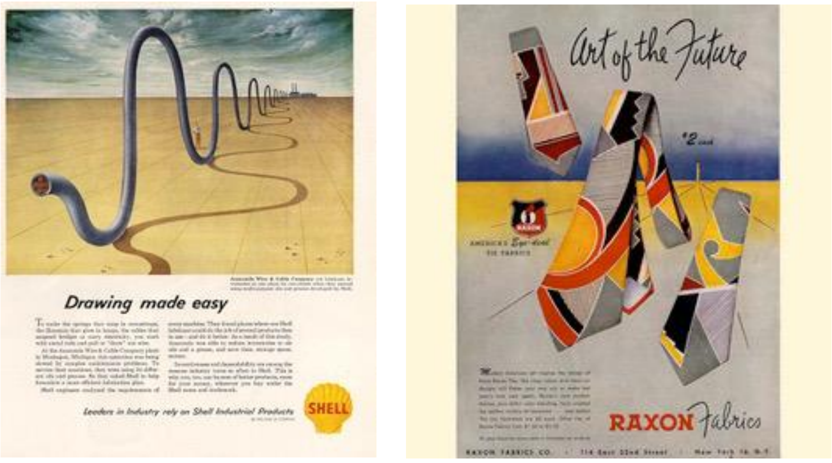

Figure 4. Two ads (the one left from 1958, the one to the right undated) that display a typical spatial solution of "modernity and progress".

The **second problem** we will deal with is how the annotation of an image has to be prepared so that the AI understands, not only what is shown, but what the image is about. We ask ourselves what formations in the picture would be optimal for the learning process of the neural networks, and believe that visual narratives, agency, and visual metaphorics are highly relevant. This is a methodology for teaching the AI relevant modes of visual literacy and cultural competence. This competence is coded through historically developed visual conventions. Thus, specialized art-historical, semiotic, and narratological concepts describing pictorial conventions will be utilized in the annotations. We will also test the optimal amount of additional textual information paired with the image.

5. Concluding Discussion

As we have seen, the development within artificial intelligence has been very rapid, and not least within image and natural language analysis. However, most of the methods used today are based on statistical predictions of the material. As such, it is difficult to talk about cognitive abilities, when it comes to understanding of external information. This is also still to some extent a sparsely researched area. There are some systems, that show a remarkable skill in using multi-modal text and image parity, for example CLIP, as we have described previously. However, as mentioned above, these systems will still not be able to transcend the border between identification and understanding. Identification is very important for many applications, but for understanding we assume that a different approach may be necessary.

Our wish is for a system architecture for specialized image retrieval where: firstly, the richness of natural language descriptions is preserved, secondly, where art-historical terms describing pictorial conventions should be utilized, thirdly, a certain amount of contextual information is included, and lastly, where the *semantic gap* can be traversed, that is, the discrepancy between the information that can be derived from the low-level image data (color, shapes) and the interpretation that human viewers of an image base on their visual literacy and cultural competence.

Acknowledgements

This project is partly funded with grants from CIRCUS (*Support for cross-cutting research projects),* which is gratefully acknowledged. Thanks also go to the University Library at Uppsala University, which has given access to the journals from which the pictures for analysis are taken.

References

[1] G. Boehm, *Wie Bilder Sinn erzeugen. Die Macht des Zeigens*. Berlin: Berlin University Press, 2007.

[2] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, "Large-Scale Visual Relationship Understanding," *AAAI*, vol. 27, 2018, Accessed: Jan. 25, 2022. Available: https://www.semanticscholar.org/paper/Large-Scale-Visual-Relationship-Understanding-Zhang-Kalantidis/847d0b91b60d8b4082d32bcbd898185c831af1d7

[3] M. Marinescu, A. Reshetnikov, and J. Moore, "Improving object detection in paintings based on time contexts", Virtual proceedings, pp. 926–932. doi: 10.1109/ICDMW51313.2020.00133.

[4] M. Springstein, S. Schneider, J. Rahnama, E. Hüllermeier, H. Kohle, and R. Ewerth, "iART: A Search Engine for Art-Historical Images to Support Research in the Humanities," in *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event China, Oct. 2021, pp. 2801–2803. doi: 10.1145/3474085.3478564.

[5] J. von Bonsdorff, "Visual Metaphors, Reinforcing Attributes, and Panofsky's Primary Level of Interpretation," in *The Locus of Meaning in Medieval Art - Iconography, Iconology and Interpreting the Visual Imagery of the Middle Ages.*, vol. 2019, L. Liepe, Ed. Berlin, Germany: Medieval Institute Publications, pp. 110–127.

[6] G. Lakoff and M. Johnson, *Metaphors We Live By*. Chicago: University of Chicago Press, 1980.

[7] G. J. Holzmann, *Beyond Photography - The Digital Darkroom*, vol. 1988. Prentice Hall.

[8] A. M. Spalter, *The Computer in the Visual Arts*, vol. 1999. Reading, Massachusetts: Addison-Wesley.

[9] J. Webber, "The Ethics/ Skills Interface in Image Manipulation," *Australasian Journal of Information Systems*, vol. 7, no. 2, Art. no. 2, 2000, doi: 10.3127/ajis.v7i2.265.

[10] A. Elgammal, M. Mazzone, B. Liu, and D. Kim, "The Shape of Art History in the Eyes of the Machine," presented at the AAAI, New Orleans USA, Feb. 2018.

[11] R. C. Schank, "Conceptual Dependency: A Theory of Natural Language Understanding", *Cognitive Psychology*, vol. 3, pp. 552–631, 1972.

[12] S. L. Lytinen, "Conceptual dependency and its descendants," *Computers & Mathematics with Applications*, vol. 23, no. 2–5, pp. 51–73, Jan. 1992, doi: 10.1016/0898-1221(92)90136-6.

[13] C. J. Fillmore, *THE CASE FOR CASE*. 1967. Accessed: Feb. 04, 2022. Available: https://eric.ed.gov/?id=ED019631

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, Accessed: Feb. 09, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[15] W. Su *et al.*, "VL-BERT: Pre-training of Generic Visual-Linguistic Representations," *arXiv:1908.08530 [cs]*, Feb. 2020, Accessed: Feb. 09, 2022. [Online]. Available: http://arxiv.org/abs/1908.08530

[16] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021, pp. 8748–8763. Accessed: Feb. 22, 2022. Available: https://proceedings.mlr.press/v139/radford21a.html

[17] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, Sep. 1980, doi: 10.1017/S0140525X00005756.