https://doi.org/10.2352/EI.2022.34.13.CVAA-170

Extracting associations and meanings of objects depicted in artworks through bi-modal deep networks

Gregory Kell,[a] Ryan-Rhys Griffiths,[b] Anthony Bourached,[c]
and David G. Stork[d]
[a] Department of Population Health, King's College London, London, UK
[b] Department of Physics, University of Cambridge, Cambridge, UK
[c] Oxia Palus, London, UK
[d] Consultant, Portola Valley, CA 94028 USA

ABSTRACT

We present a novel bi-modal system based on deep networks to address the problem of learning associations and simple meanings of objects depicted in "authored" images, such as fine art paintings and drawings. Our overall system processes both the images and associated texts in order to learn associations between images of individual objects, their identities and the abstract meanings they signify. Unlike past deep net that *describe* depicted objects and infer predicates, our system identifies meaning-bearing objects ("signifiers") and their associations ("signifieds") as well as basic overall meanings for target artworks. Our system had precision of 48% and recall of 78% with an F1 metric of 0.6 on a curated set of Dutch *vanitas* paintings, a genre celebrated for its concentration on conveying a meaning of great import at the time of their execution. We developed and tested our system on fine art paintings but our general methods can be applied to other authored images.

1. INTRODUCTION: EXTRACTING MEANINGS CONVEYED THROUGH IMAGES

Despite recent successes in semantic image analysis,[1,2] a large class of challenging image analysis problems have yet to be addressed by automatic methods, specifically inferring *meanings* expressed in images. While traditional semantic image analysis can learn and represent the identities of physical objects, their relations, and general context (indoor/outdoor, day/night), it has yet to address adequately the problem of inferring the motives and messages of an image's creator or "author."

The general problem of extracting a plausible meaning from an arbitrary authored image is surely AI complete. Specifically, such meanings are context dependent, and viewers bring a wealth of general commonsense knowledge and domain-specific information when interpreting such images. This context information is generally represented in automatic systems somewhat abstractly and symbolically, ultimately derived from texts. It is for such reasons that our system, described in Sect. 4, relies on both natural language processing *and* image analysis; it also relies on a bi-modal knowledge representation that facilitates simple inference based on both image and symbolic or textual knowledge.

The overwhelming majority of natural photographs used in semantic image analysis record a scene but carry few if any specific messages beyond the obvious, such as (I LIKE THIS or I WAS HERE). By contrast, that vast majority of Western artworks were crafted to convey a message, moral, or *meaning*, often at patrons' requests. For example, the Catholic Church, as patron, commissioned numerous artists to depict stories, episodes and messages from the Bible, to adorn religious sites and documents. Such messages ranged from specific, such as ADAM AND EVE DISOBEYED AN ORDER AND COMMITTED MANKIND'S ORIGINAL SIN, or more ones that are more general, such as HAVE COMPASSION FOR THOSE LESS FORTUNATE. Likewise, Hindu artworks convey morals such as WITHOUT BEING ATTACHED TO THE FRUITS OF ACTIVITIES, ONE SHOULD ACT AS A MATTER OF DUTY; in Greek mythology ZEUS CONTROLS YOUR FATE; in political art COMMUNISM IS EVIL; and in advertising BUY COCA-COLA. No photograph—natural or "art" photograph—has a complexity and depth of meaning as some masterpiece paintings, such as Diego Velàzquez's *Las meninas*, which is the most analysed painting in all of art.[3]

A key challenge to AI is that any one of these messages might be conveyed through innumerable images that at least on the surface bear little relation to one another. One need merely think of all the images in a

print advertising campaign for life insurance policies to appreciate such a breadth. While there may be multiple plausible readings of a work, some readings are consistent with more information, or the "ground truth" of an author's stated goals. Below we shall address this problem in a restricted domain: the *vanitas* artworks of the Dutch Golden Age. These works were created with clear intention to convey religious and moral messages. Note, though, that we address this art genre because the problem domain is well circumscribed—in the objects that are depicted, the styles, and messages. We stress the general problem is broader than just the analysis of this genre: other artwork, commercial advertisements, political propaganda, and some information graphics convey meaning or import that may be amenable to our methods.

Our work presented here is inspired by recent work using deep networks that address a related problem: identifying key figures or "actors" in paintings as a step to understanding a work's meaning.[4] That previous approach automatically identified and located key signs or iconographic *attributes* in an artwork, and found the closest segmented human figure. Thus, for instance, a segmented human figure would be automatically identified as St. John the Baptist because of its proximity to his iconographic attribute, the crucifixion cross; likewise Christ was identified by his proximity to one of his attributes, a dove; and so on.

Semantic analysis of images involves relating images to concepts expressed in text. In that prior work, the representation and data structure linking images with text consisted of a list of associations of semiotic attributes with actors (specifically saints depicted in paintings)—a relation that was entered by hand. In the system reported in the present paper, such an association was learned through natural language processing of texts associated with art images.

2. SEMIOTICS: SIGNIFIERS AND SIGNFIEDS

The nature, representation, and reasoning related to meaning is central to the branch of philosophy known as epistemology, to the field of semiotics, and to much of cognitive science and artificial intelligence. Most work in AI on inferring meanings has been restricted to documents and natural-language text, in large part because almost all texts were written by authors intending to convey information and non-trivial meaning, while the same cannot be said for the majority of natural photographers.[5] Computational approaches to visual semiotics are not as well developed as the corresponding approaches to text analysis, but has nevertheless shed light on a number of problems.[6]

The essential items of concern in semiology are *signifiers* and *signifieds*. Meaning derives from relations and associations between certain objects, marks, or works—signifiers—and their associated objects, concepts, or ideas—signifieds. There are three basic classes of signifiers:

Signs In principle, signs can be anything, so long as there is some conceptual link from them to a signified. For instance a smashed wineglass on the floor may refer to the (absent) person who threw it there.

Icons Icons resemble their associated signifieds but are simpler and more abstract. For instance the digital pixel-based icons of printers and file folders on computer desktops refer to physical printers and folders and, indirectly, their function.

Symbols Symbols are arbitrary conventions, often created by stakeholders as, for instance **$**, which indicates DOLLARS, or more generally MONEY.

All three classes of signifiers appear in some paintings and support the creation of meaning of such works.

2.1 Herman van Steenwijk's *Still life: An allegory of the vanities of human life*

Consider an example of how signifiers and signifieds work in Herman van Steenwijk's *Still life: An allegory of the vanities of human life*. The most compelling message of the work is: DO NOT CONCERN YOURSELF WITH LIFE OF THIS WORLD, BUT INSTEAD PREPARE FOR ETERNAL LIFE TO COME. Note that all educated viewers (and most illiterate viewers) in 17th-century Netherlands would understand this message, which was also preached from pulpits, widely written and discussed. Indeed, this message was associated with the Dutch Revolt and ultimate break from its rulers in Catholic Spain.

Figure 1. Herman van Steenwijk's *Still life: An allegory of the vanities of human life* (c. 1640). Nearly every object in this work is a sign and contributes to the overall message and meaning of the work, which is fairly explicit in the work's subtitle. Moreover its highly realistic style and composition (eg., with light shaft directed toward the skull) support the work's primary meaning.

How is that message conveyed, and how might an automated system extract or infer such a message?

The skull is a sign that refers to mortality and the inevitability of death; the books are signs that refer to worldly (not otherworldly) knowledge; the musical instruments—flute, sackbut (early trombone) and lute (early guitar)—are signs that refer to culture and sensual pleasures; the lovely seashell and samurai sword refer to worldly travel and luxury; the open pocket watch refers to the passage of time and human life; the oil lamp at the rear refers to the passage of time and life and the fact that it has recently been extinguished (as evident from the thin trail of smoke) refers to the fact that life can end at any moment. Finally, the shaft of light directed toward the skull has religious connotations as well.

Our overall technical goal is to compute simple meanings given such a vanitas painting, specifically the references or significations of objects and relations withing the painting to simple meanings. Our computational approach is to *learn* these signifier-signified relations from existing texts describing such artworks and use that information as part of a system that identifies these meanings of objects within the artworks. Specifically, our system will learn relations such as SKULL:MORTALITY, BOOK:WORLDLY KNOWLEDGE, and so on, as can be represented in a knowledge graph, as shown in Fig. 13. Further, computer vision and pattern recognition will identify the signifiers from the image alone. To the best of our knowledge, the work we present here is the first to use natural language processing and computer vision to extract simple meanings from authored images.

3. DESCRIPTION OF TASK AND MODELS

Our overall task is to develop a system that takes an art image as input and performs simple interpretation of its meaning, particularly the meaning of depicted objects. Thus, for instance, such a system would allow a user to point to an object in a painting and ask "What does that mean?" and the system would not only provide the identity of an object but also its simple meaning, for example THAT IS A SKULL, WHICH REPRESENTS MORTALITY.

3.1 System components

In order to compute the below interpretation of the work in Fig. 2, the model would need to not only detect the objects in the image, but also associate them to the relevant concepts. For instance, the lute, music and inkstand

Figure 2. Adam Bernaert's *Still Life* (1665).

Figure 3. *Vanitas Still Life J Falk (Dutch, 1600-1699)* The symbolism of the skulls in this painting is obvious, but the rose (quick to wilt) and oil lamp (easily snuffed out) also refer to life's brevity and fragility. The vanitas symbolism is underscored by the Latin inscription underneath: "All that is human is smoke, show, vanity and the picture of a stage."

would need to be recognised and associated with the creative endeavours by the same system. Thus, the system requires an object recognition component that detects the signifiers and a knowledge graph that associates the signifiers with the signifieds. The knowledge graph can be automatically created by leveraging natural processing techniques to extract information from texts describing the symbolism of vanitas artworks and using the relevant entities and relations to create a knowledge graph.

The accompanying text describes the work in Fig. 2:

> These objects symbolize transitory human achievement and satisfactions. The atlas is open to a map of the East Indies, source of many Dutch fortunes, and there is a city council document with an imposing seal. The other open book is a history of the early counts of Holland-whose lands were absorbed by the dukes of Burgundy in the 1400s. The lute, music, and inkstand represent creative endeavors, which, like satisfaction in beautiful objects such as pearls, are transitory pleasures. Even the heavens and the earth, represented by two globes, are effected by Time, whose relentless passage is marked by the hourglass.

4. SYSTEM ARCHITECTURE

Our system architecture, shown in Fig. 12, consists of image analysis and natural language processing.

Figure 13 shows an example of a knowledge graph that could be produced using the natural language processing (NLP) techniques described below.

Figure 4. *Still Life, 1643 Pieter Claesz.* Everything on the table, from the fluted glass and goblet to the lobster and crab, is indeed life-like. You can almost smell the lemons. The Dutch proudly displayed such expensive status symbols in their homes, the exotic food and material possessions reminding them of the good things in life, even as the watch reminds them of their transience. The bread and wine, in a touch of Christian symbolism, echo the moralizing message of vanitas, or vanity: all earthly things must pass.

4.1 Visual processing

A key component of our system is a deep convolutional network object identification and localization module. We take an implementation of mask R-CNN[7,8] that has been pre-trained on the Microsoft COCO (Common Objects in Context) dataset[9] and apply transfer learning to our dataset of Vanitas images.

4.2 Natural language processing

The objects detected by the vision component can be mapped to abstract meanings using a knowledge graph. An example of one that was generated automatically is shown in Fig. 13.

5. TRAINING AND TESTING

5.1 Natural language processing

The knowledge graph was created using texts from pages on the internet. This involved retrieving the URLs of the websites from Google using the query "vanitas meaning" and scraping the corresponding websites. The total number of pages scraped was 96.

The texts containing references to the artists and paintings included in the test set for the vision modules was reserved for evaluating the knowledge graph (27 pages). The rest of the texts were used as the train set for constructing the graph (69 pages).

5.2 Training

5.3 Natural language processing

The object-meaning pairings were extracted by performing semantic role labelling on the texts using a BERT-based model.[10] The heads (objects) of the knowledge graph were set to be the proto-agents (Arg0) or if none were available, the proto-patients (Arg1). All the other arguments were assigned as tails (meanings) of each corresponding head (given a predicate). Each edge was weighted by the number of times a head-tail paring appeared in the texts, i.e., an appearance of 2 corresponded to a weight of 2. The knowledge graph was pruned by including only the heads that represented the objects detected by the vision module. The tails that corresponded to undesirable entities, i.e., person, organisation, geopolitical entity, location, were excluded. The transition-based entity recognizer provided by the Spacy library was used for this purpose. Finally, any edges with a weight less than 2 were removed.

Figure 5. *1642, Vanitas Still Life with Flowers and Skull, Adriaen van Utrecht* In his 1642 painting, Vanitas Still Life with Flowers and Skull, Adriaen van Utrecht depicts a multitude of objects, including but not limited to a vase of flowers, a human skull, small gold and silver coins, two glass vases, and a book. In the tradition of still-life painting, these objects have individual meanings all their own. For example, the orange book beneath the skull symbolizes human knowledge. Oftentimes in still-life paintings artists would choose to paint books with their pages open, exposing their content to the viewer as a window into the intellectual realm (note 1). In front of the book, draping softly over the edge of the table, are several gold and silver coins, representations of wealth and status. Another manifestation of the vanitas theme can be found in the timepiece that rests next to the scintillating currency. The cover of this gold clock remains open for the viewer to tell the time and be reminded, once again, about the inevitable passing of time. Although integral to the composition, these symbolic objects do not expose the intrinsic meaning of Utrecht's still life. Indeed, this painting's meaning transcends these objects' symbolism, and it can arguably be found in relationship forged between the bouquet and the skull.

The light radiating from the bouquet of flowers on the left side of the canvas serves as a point of entry into Utrecht's cluttered, frenetic composition. Of the flower varietals represented in the glass vase, the bright white and yellow flowers are first to captivate the viewer's attention. Their vibrancy refuses to be ignored, and light seems to glow from the depths of their petals. The flowers found at the base of the glass vase seem to have withered away and fallen from the cohesive bouquet above. One of the rose buds, along with the metal chains and a scrap of paper, hangs limp over the table's edge, boldly entering the viewer's space and creating a trompe l'oeil effect. Most of the light in this painting falls on the bouquet and the skull, likely an intentional way for Utrecht to highlight the two most important objects of the painting. Prior to becoming prominent elements within still-life paintings, skulls were oftentimes painted on the back of portraits to remind their owners of life's brevity (note 2). Placing a flower arrangement directly next to a human skull was surely intentional on the part of the artist, and their proximity forces a dialogue about the relationship between life and death. A craggy, decaying skull in the presence of vibrant, budding flowers seems unsettling at first, and initially the viewer cannot help but wonder how the two could ever be related. A closer look, however, reveals that a profound connection does, indeed, exist between the two ostensibly polar opposites. As the title indicates, Utrecht's painting is replete with evidence of the theme of vanitas, a theme that tenderly reminds us about the transience of life. Skulls are particularly interesting in that they are no longer part of a living human, nor are they really an inanimate object (note 3). They fall somewhere in the middle, emphasizing their power as symbols of vanitas and reminders to cherish life before death. Utrecht's work calls the viewer beyond vanitas, however. More than merely symbols of knowledge, life and death, these objects invite the viewer to contemplate the culture of cultivation in which they once existed.

Figure 6. *Vanitas Still Life, 1603, Jacques de Gheyn II* De Gheyn was a wealthy amateur who is best known as a brilliant draftsman, but he also painted and engraved. This panel is generally considered to be the earliest known independent still-life painting of a vanitas subject, or symbolic depiction of human vanity. The skull, large bubble, cut flowers, and smoking urn refer to the brevity of life, while images floating in the bubble—such as a wheel of torture and a leper's rattle—refer to human folly. The figures flanking the arch above are Democritus and Heraclitus, the laughing and weeping philosophers of ancient Greece.

Figure 7. *Still Life with a Skull and a Writing Quill, 1628, Pieter Claesz* In this still life, close observation and realistic detail operate in tension with explicit symbolism. The toppled glass, gap-toothed skull, and guttering wick of an oil lamp all serve as stark symbols of life's brevity. Working with a limited palette of grays and browns, Claesz carefully describes the surfaces of these unsettling objects. By arranging them on a pitted stone ledge, the artist connects the picture's space to our own, making the message all the more compelling.

Figure 8. *Vanitas Still-Life Hendrick Andriessen, ca 1650* Before you read further, take a few moments to just look at this remarkably detailed painting. As your eyes move across the canvas, try to identify the objects that you see on the table. Often when we think of a still-life, we imagine a painting of fruit or flowers, so you might be surprised by some of the objects you see. This is a vanitas—a specific type of still-life that emerged in the 17th century in the Netherlands and grew out of a long artistic tradition known as memento mori, meaning "reminders of mortality." While looking closely at this painting, you probably noticed several objects that could be called reminders of mortality, such as the skull, the wilting tulip, and the dying wick of the candle.

Vanitas still-lifes were appreciated for both their beauty, rendered in incredible detail, and for their deeper symbolic significance. Andriessen's contemporary audience may have recognized the crown as a specific, haunting reference to the recent execution of King Charles I of England in 1649. Every element of this painting also has broad symbolic power: the skull, bubbles, extinguished candle, flowers, and glass vase remind the viewer of the impermanence of life; the watch symbolizes the passing of time; the jeweled crown and bishop's mitre lying behind it point to the fleeting nature of power.

Figure 9. *Vanitas Still Life Unknown artist, possibly Flemish After 1649* The subject here refers to Charles I's troubled life and is a testament to the suddenness of death and the vanity of early power and glory. For example, the bubbles pertain to the brevity of Charles I's life (he was beheaded at the age of 44), the broken skull conveys the fragility of human beings, and the globe symbolizes the power and possessions that death steals away.

Figure 10. *Heem, Jan Davidsz. de, Still-Life, c. 1630* This picture, showing a skull, a book and roses, is a memento mori, presenting a sinister contrast between the skull with its empty sockets and the fragrant pink rose so full of life.

Figure 11. *Gheyn, Jacob de II, Vanitas Still-Life, 1603* The dominant motifs in the picture are a human skull and, floating above it, a transparent sphere or bubble. These forms occupy a stone niche with a slightly pointed arch, the keystone of which is inscribed HVMANA VANA (Human Vanity). The spandrels flanking the arch are filled with sculptural figures of philosophers with books at their feet? to the left, Democritus, who gestures toward the globe and laughs; and, to the right, Heraclitus, who points to the sphere and weeps. The sphere purposefully resembles a soap bubble, the familiar vanitas motif. Two common vanitas symbols, cut flowers and smoke, rise from urns at either side of the niche. The coins depicted at the bottom of the composition on the sill between the vases were used as currency in the Netherlands about 1600. One of them, the silver medal of 1602 commemorates the capture of a Portuguese galleon by two Zeeland merchant ships earlier that year, off Saint Helena in the South Atlantic.

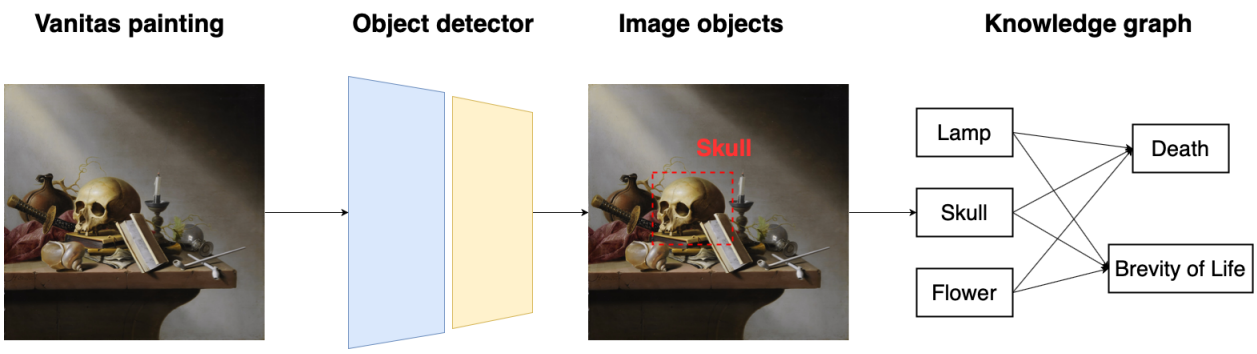

Figure 12. System architecture. The painting is provided as input, which is then processed to recognize and locate the objects. Each recognized object is then passed to the knowledge graph, which links that object (signifier) to its meaning (signified).

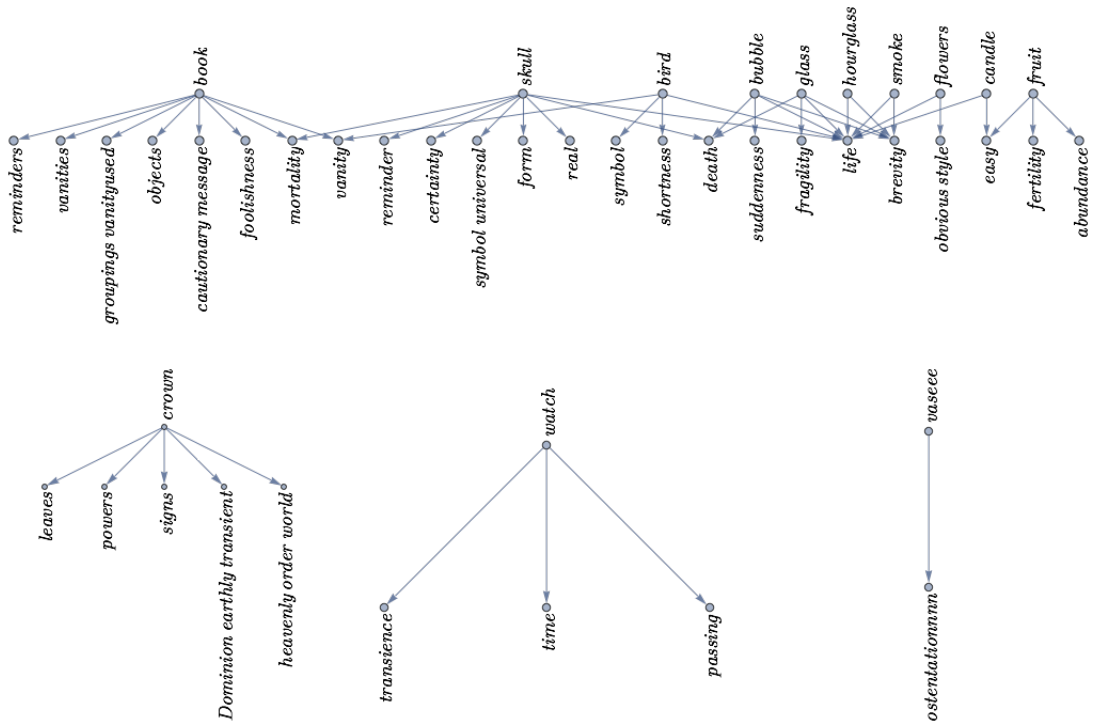

Figure 13. Our bipartite knowledge graph linking signifiers to signifieds, learned from textual descriptions of vanitas paintings written by contemporaneous critics and art historians.

5.4 Visual processing

We transfer learn the weights of the pre-trained mask R-CNN model on 56 images of vanitas paintings. The training procedure operates in two phases: First, we freeze the weights of all layers save for the final layer of mask R-CNN and fine-tune the weights of this final layer for 10 epochs using a learning rate of 0.001. Second, we fine-tune all layers of the network for 2 epochs using a learning rate of 0.0001. Training is performed on a single NVIDIA Tesla K80 GPU in the Google Colab environment.

5.5 Testing

5.6 Natural language processing

The 27 texts in the test set were used to extract the object-meanings pairings manually. This led to the creation of a subset that matched the objects and meanings from descriptions in specific paintings and of a another consisting of more general pairings (i.e., what associations between objects and meanings existed for vanitas paintings in general). When evaluating the knowledge graph exact matches and partial matches between the meanings of the knowledge graph and the test set were used to calculate the precision, recall and F1 score. An exact match was defined as the instance where a meaning in the knowledge graph (i.e., 'life') and the meaning in the test set (i.e., 'life') were identical. A partial match was when a meaning in the knowledge graph (i.e., 'life') was a subset of the meaning in the test set (i.e., 'brevity of life').

Additionally, the semantic similarity between the meanings in the knowledge and the test set was used to define matches (true positives), i.e., if the cosine distance between the BERT sentence embeddings[11] was above a predefined threshold (0.7) (i.e. 'time' and 'transience'). If the similarity between a meaning in the knowledge and all of the meanings in the test set for a given object was below the threshold, the meaning in the knowledge graph was deemed a false positive. Likewise, if the similarity between a meaning in the test and all the meanings in the knowledge for a given object was below the threshold, the meaning in the test set was deemed a false negative.

5.7 Visual processing

We test the mask R-CNN object detector module on 18 heldout images of Vanitas paintings. As with training, inference is again performed on a single NVIDIA Tesla K80 GPU in the Google Colab environment. All code to reproduce our results is made available at https://github.com/gck25/fine_art_asssociations_meanings.

5.8 End-to-end system

The full system was tested on 8 images from the test set for which descriptions relating to each specific image could be found in the web texts. The system's ability to map paintings to meanings was assessed. The same metrics used to evaluate the NLP component was used to evaluate the end-to-end system.

6. RESULTS

We express our results using *precision* and *recall*, where

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN} \tag{1}$$

and TP are true positives, FP are false positives, and FN are false negatives.

6.1 Natural language module and processing

The semantic and partial matches shown in Table 1 are significantly higher than the exact matches. This suggests that although the knowledge graph does not associate objects with exact phrases in the test set, it does connect them to meanings that are semantically similar.

6.2 Vision module and processing

According to Table 1, the precision is substantially larger than the recall. The majority of the tags assigned by the component are correct, but it is missing tags for a large number of objects.

Table 1. Precision, recall, and F1 performance for each component in the overall system.

Component	Precision	Recall	F1
Knowledge Graph (exact)	0.11	0.04	0.06
Knowledge Graph (partial)	0.26	0.18	0.22
Knowledge Graph (semantic)	0.49	0.39	0.43
Object Detector	0.6	0.06	0.10
End-to-end System (exact)	0.04	0.11	0.06
End-to-end System (partial)	0.12	0.33	0.17
End-to-end System (semantic)	0.48	0.78	0.60

6.3 End-to-end system

The end-to-end system follows the same same trend as the knowledge graph in terms of the relative difference in the metrics between exact, partial and semantic matches. Nonetheless, the recall is larger than the precision. This could be explained by the difference in test sets used to evaluate each component. The knowledge graph test set was created using texts that described Vanitas symbolism in general, while the end-to-end test set used texts referring to specific images. The painting-specific texts associate a lower number of meanings to each image, which increase the recall and lower the precision.

7. CONCLUSIONS AND FUTURE DIRECTIONS

We have demonstrated computational approaches to the extraction of simple meanings in authored images, specifically vanitas paintings from the Dutch Golden Age. Our method involves natural language processing of contemporaneous and scholarly texts commenting on such artworks, to build a bipartite knowledge graph relating visual signs in the paintings to signifiers. Our end-to-end system had precision and recall of 0.48 and 0.78, respectively. Our work may serve as a base for higher-level semantic interpretations of images, including additional genres, such as religious art, mythological art, print advertisements, and political propaganda.

Our work is a very early—yet, we feel, promising—step in a broad program of understanding author-created meanings in images. Art is a rich source of problems in this domain and far richer than the domain of natural photographs, which dominates research in computer vision. This nascent research program expands the class of problems faced by AI and, once enhanced, may serve as a tool for art scholars.[12]

Acknowlegements

The last author would like to thank the Getty Research Institute, Los Angeles, where some of the work on this project was performed.

Code available

Our code is available at: `https://github.com/gck25/fine_art_asssociations_meanings`.

REFERENCES

1. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077–6086, 2018.
2. V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Applications* **39**(12), pp. 2481–2495, 2017.
3. S. L. Stanton-Pruitt, *Velàquez's 'Las meninas'*, Cambridge University Press, Cambridge, UK, 2002.
4. D. G. Stork, G. H. Cann, A. Bourached, and R.-R. Griffiths, "Computational identification of significant actors in paintings through symbols and attributes," in *Computer vision and analysis of art*, D. G. Stork and K. Heumiller, eds., IS&T, (San Francisco, CA), 2021.

5. V. Targon, "Toward semiotic artificial intelligence," *Procedia Computer Science* **145**, pp. 555–563, 2018.
6. G. Aiello, "Visual semiotics: Key concepts and new directions," in *SAGE handbook of visual research methods*, L. Pauwels and D. Mannay, eds., ch. 23, pp. 367–380, SAGE Publications, Thousand Oaks, CA, 2019.
7. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
8. W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow." `https://github.com/matterport/Mask_RCNN`, 2017.
9. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.
10. P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," *CoRR* **abs/1904.05255**, 2019.
11. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *CoRR* **abs/1908.10084**, 2019.
12. D. G. Stork, "Automatic extraction of meaning in authored images such as artworks: A grand challenge for AI," *ACM Transactions on Cultural History Computing* **3**, 2022 (in press).