VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wānanga o te Ūpoko o te Ika a Māui

School of Engineering and Computer Science
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Internet: office@ecs.vuw.ac.nz

Photo2Sketch Project Proposal

Leyton Blackler

Supervisor: Fanglue Zhang

Submitted in partial fulfilment of the requirements for
Bachelor of Engineering with Honours - BE(Hons).

Abstract

This project proposal outlines the idea of converting two dimensional photographic images into simplistic line drawing representations in the style reflective of that of a human. A series of deep machine learning techniques will be used to achieve such a system, primarily utilising the concepts of salience detection, instance segmentation and generative adversarial network image synthesis.

1. The Problem

Identifying subjects within two dimensional images depends upon the ability to distinguish objects from their surroundings. Often photographic images contain detail that distracts from the main subject of the image. This detail may only exist to add additional complexion and texture to the subject, or may exist as part of the surroundings of the subject. It is often still possible that the image could exist without this detail, while still clearly portraying the original subject. Inessential detail within an image can add an element of difficulty when attempting to identify the subjects, both in human vision as well as in computational image processing and means a higher amount of image data is required in order to communicate the information that the image represents. Simplifying and improving the ability to identify and communicate subjects of a two dimensional image can lead to a number of beneficial applications.

Line drawings are effective at portraying a subject with very little detail. The most basic form of line drawings are typically black and white, simply utilising a series of plain lines to abstract an image into a form where only the most important elements of the subject are preserved. Humans have consistently and continuously utilised this simple visual form of communication for centuries in order to document and transfer ideas and knowledge, dating back to primitive cave drawings or ancient hieroglyphics. It is evident that basic human created line drawings are an effective form of communication. Because of these features of line drawings, automatic and accurate conversion of a two dimensional photographic image into line drawing representations would allow for simplified image subject identification and communication of the subject of an image. An example application area where line drawing generation may be of benefit is in relation to the research project *Sketch2Photo: Internet Image Montage* explored by Tao Chen et al [1]. This project demonstrates how basic line drawings can be utilised in combination with photomontage techniques for realistic image synthesis. This technology, in conjunction with the ability to generate line drawings from existing photographic images, would mean that a photographic image could be used to synthesise an entirely new photographic image with the same subject positioning and identities. This would be useful for a wide variety of instances. One such application may be if there were a photographic image with the correct subject identities and positioning, but a different perspective or style of image was desired.

Existing solutions for two dimensional photographic image to line drawing conversion are unable to create satisfactory results. Previous related work demonstrates methods of creating line drawings relying primarily on the use of edge detection. While this achieves an acceptable extraction of the structural lines of the subject, details surrounding the subject as well as some inessential details that make up the subject itself still remain. Additionally, these line drawings are created from the photographic depiction of the actual object. Typically when a human naturally creates a line drawing of an object, the depiction is in a form more simplified than the realistic counterpart. While the drawing would follow the general poise and shape of the object, there may be slight augmentations in the exact edge sizes and positions. Therefore, this means an edge-based line drawing resembling the original photo still contains additional inessential detail and would not sufficiently resemble a line drawing drawn by a human. An example of this, is shown by the research paper *Coherent Line Drawing* by Henry Kang et al [2]. Figure 1 shows an original photograph that was used as the input image on their proposed system. Figure 2 demonstrates the output of the system. While still technically consisting of only a series of lines, the resulting line-based image contains significant amounts of detail from the surroundings of the subject, as well as a large amount of detail comprising the subject itself. Figure 3 is an example of a human line drawing of the same butterfly as obtained from the Sketchy database of image to sketch pairs [3].

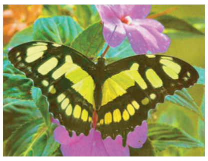

Figure 1: Original photo

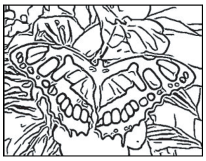

Figure 2: Edge detection line based representation

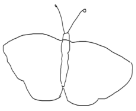

Figure 3: Human line drawing

Based upon this, it is evident that there are differences between edge line extraction line representations and representations drawn by humans. When humans draw a subject by means of a basic line drawing, a significant portion of the original detail is redacted and the resulting image is still able to portray the subject in a much more concise manner. Therefore, the aim of this project is to implement and evaluate a software system that is able to automatically convert a two dimensional photographic image into a high quality, yet basic line drawing that simplistically and clearly portrays the subject of the original image in the same style that would be achieved if drawn by a human.

2. Proposed Solution

Due to the varying nature of the exact edge sizes, shapes and positions of a line drawing compared to edge detection of the original photograph, the solution I propose will involve synthesis of a new image using generative adversarial network technologies. A generative adversarial network is a framework for machine learning systems where two sub-networks, the generator network and discriminator network, contest each other in order to both improve. The generator network is responsible for generating new samples, in the case of this project, the generator will produce line drawings. The discriminator attempts to determine whether the generated line drawing was generated by the generator, or is a real line drawing created by a human. Through each iteration, each of the sub-networks improve; the generator becomes better at creating realistic line drawings, and the discriminator becomes better at determining whether a line drawing was drawn by a real human, or the generator. For this, I plan to utilise the Sketchy Database for training purposes [3]. The Sketchy Database contains 75,471 human-drawn line drawings of real photographs of 12,500 objects across 125 categories.

In order to detect the important subjects of the photographic images, my proposal also intends to incorporate salience detection. This means that the system will be able to determine which the area of a given photographic image is considered the most important area of the image. The most important area however, does not specifically indicate that the area is representative of a specific object. Therefore, I additionally propose to include the functionality of image instance segmentation in order to purely define the core subject (or subjects) of an image. This means that if for example, a photographic image contains two butterflies, the generative adversarial network will not attempt to draw the butterflies collectively, but will instead acknowledge that each is a separate subject that should be drawn individually.

To achieve this, I aim to explore existing research that aligns closely to the ideas that will be required to reach my solution. By using existing systems, I will be able to incorporate some of the existing knowledge discovered by this research to better advantage the likelihood of successful results from my solution. This may involve devising my own implementations of certain models and algorithms, or building upon and adapting existing implementations for certain components of my solution.

This solution I propose in order to implement such a system involves utilisation of cutting edge deep learning techniques. The primary machine learning framework I plan to use in order to perform deep learning operations and model training is TensorFlow. However, I am also open to the possibility of exploring alternative machine learning frameworks, such as PyTorch, if it is discovered that TensorFlow is unsuitable.

The project timeline is constrained to the boundaries imposed by the Victoria University of Wellington 2019 academic year, with the exclusion of the third semester. Therefore, the project timeline commences in week one of the first semester in March, and will conclude at the beginning of the examination period of the second semester in October. I plan to manage this project with an agile approach. I believe this will be most effective due to the initial lack of knowledge regarding how I will exactly implement such a solution and the phases of development that my solution will entail. This will allow me to freely and consecutively research and iterate system designs throughout the project timeline.

3. Evaluating your Solution

The solution that I propose to implement will be a form of pipeline in which an input image will pass through in order to be converted into a line drawing. Throughout this process, there will be several main steps in the pipeline that can each be evaluated individually to assess the performance of that particular area of the system. These components include the areas of salience detection, instance segmentation and then the image synthesis via the generative adversarial network. Together, the performance of these components that create the overall system can be evaluated to give an effectiveness of the system as a whole.

Salience detection can be evaluated using the salience ground truth maps of original photographic images. This allows for a generated salience map of a photographic image to be compared against the ground truth map to determine how accurate the system was able to detect salience. An example of this is shown in the research project *Review of Visual Saliency Detection with Comprehensive Information* undertaken by Runmin Cong et al [4]. Figure 4 shows the original photographic image, figure 5 shows the ground truth that indicates the optimal area of salience detection, and figure 6 shows the salience detection map created using the ACSD method.

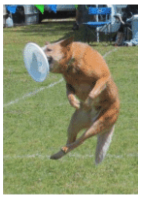

Figure 4: Original photo

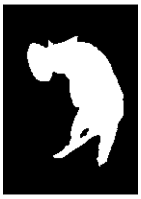

Figure 5: Ground truth

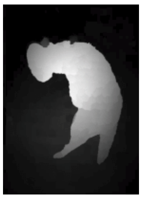

Figure 6: Salience detection

For evaluation of instance segmentation, this can be achieved in a similar fashion to how salience detection can be evaluated. For instance segmentation however, a supplementary ground truth image map with bounding boxes of the instances in the image can be compared to the instances detected by the system. This allows to measure how accurately the system is able to distinguish the individual instances of subjects within the photographic image.

For generation of the line drawings, there will be two mains areas to evaluate; the performance of the generator network and the performance of the discriminator network. The performance can be measured by determining how well the generator is able to synthesise line drawings that the discriminator cannot determine whether is real or generated, and how well the discriminator is at deciding whether an image has been generated or drawn by a human.

Due to the creative nature involved with humans drawing line drawings, there will also be an element of human judgement as to whether a synthesised line drawing is deemed to be similar to a human drawn counterpart. For this, I plan to partially utilise my own discrepancy as well as possible judgement from a test group to reduce my own bias.

4. Resource Requirements

Training models through deep learning can often be heavily resource intensive in terms of computational power. Using only a single mid-range computer could cause the process to take a significantly longer time than if a more powerful machine, or network of machines, were to be used. This could result in having to wait days or even weeks for a single iteration of a model to be trained, compared to only minutes or a few hours. Due to the nature of the project being time restrained, delays this significant would heavily hinder progress and limit the success of the project. Therefore, access to hardware offering a high level of computational power is essential.

Bibliography

[1] Chen, Tao, et al. ”Sketch2photo: Internet Image Montage.” *ACM transactions on graphics (TOG)*. Vol. 28. No. 5. ACM, 2009.

[2] Kang, Henry, Seungyong Lee, and Charles K. Chui. ”Coherent line drawing.” *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*. ACM, 2007.

[3] Sangkloy, Patsorn, et al. ”The sketchy database: learning to retrieve badly drawn bunnies.” *ACM Transactions on Graphics (TOG) 35.4* (2016): 119.

[4] Cong, Runmin, et al. ”Review of visual saliency detection with comprehensive information.” *IEEE Transactions on Circuits and Systems for Video Technology* (2018).