https://doi.org/10.2352/ISSN.2470-1173.2019.8.IMAWM-412

Frame Detection for Photos of Online Fashion Items*

Litao Hu, Jan Allebach; Purdue University; West Lafayette, IN
Gautam Glowala, Sathya Sundaram, Perry Lee; Poshmark Inc.; Redwood City, CA

Abstract

In the competitive online fashion market place, it is common for sellers to add artificial elements to their product images, with the hope to improve the aesthetic quality of their products. Among the numerous types of artificial elements, we focus on detecting artificial frames in fashion images in this paper and we propose a novel algorithm based on traditional image processing techniques for this purpose. On the other hand, even though deep learning methods have been very powerful and effective in many image processing tasks in recent years, they do have their drawbacks in some cases, rendering them ineffective compared to our method for this particular task. Experimental results on 1000 testing images show that our algorithm has comparable performance with some of the state-of-the-art deep learning models that have been used for classification.

Introduction

With the increasing popularity of numerous photo editing softwares on mobile devices and PCs, images with post-editing components are very common. This is especially the case in the online fashion market place. To make their fashion products more appealing in the competitive online market place, sellers often add artificial decorative elements in their product images, resulting in the prevalence of fashion images with non-authentic elements. With the numerous types of artificial elements, it is often challenging to effectively weed out these images from the sea of fashion images.

To make it clear, we define image frames in this paper as marginal image components that do not contain information that is useful or related to the main content in the image. An obvious example is the padding pixels around an image similar to Figure 1a. Based on the structural characteristics and locations of various images frames, fashion images frames can be divided into 7 different categories, namely horizontal frame (Figure 1a), vertical frame (Figure 1b), surrounding frame (Figure 1c), asymmetric frame (Figure 1d), frame with extra information like logos or texts (Figure 1e) and collage frame (Figure 1f). Although some frames can indeed enhance the aesthetic quality of fashion images, in most cases these frames are unnecessary and can interfere with the customer's appreciation of the main contents in fashion images.

In our proposed method, a 3-channel color input image will first be converted to a single channel gray-scale image. Then edge detection will be performed to extract edges from the gray-scale image. After that, frame detection and localization will be performed based on the extracted edges. If a frame component is not detected, the algorithm will proceed to remove any connected components corresponding to possible texts and logos from the extracted edges before performing frame detection and localization one more time. An example of a typical image frame detection process is shown in Figure 2.

*Research supported by Poshmark, Inc., Redwood City, CA 94065

(a) Horizontal Frame (b) Vertical Frame (c) Surrounding Frame

(d) Asymmetric Frame (e) Informative Frame (f) Collage Frame

Figure 1: Frame categories and examples

Figure 2: From left to right: (a) Input image; (b) Edge image; (c) Frame detection and localization result

In the following sections, we will mention some of the related work on detection of image frames. Then, we will describe in detail our frame detection and localization algorithm and its key building blocks. Experimental results comparing our methods with some deep learning models will be presented before a brief conclusion.

Related Work

Frame Detection and Localization

Digital image component analysis has been a very active research area in computer vision and image processing, especially with the increasing popularity of deep learning in recent years. There are a great variety of topics on digital image component analysis, and some most common ones are image classification[1], image segmentation, and image regions of interest detection and localization. As a sub-area under image component analysis, image frame detection also has some active research projects in recent years. In a project conducted by Allegro Tech [2], they adopted a similar rule-based frame detection algorithm

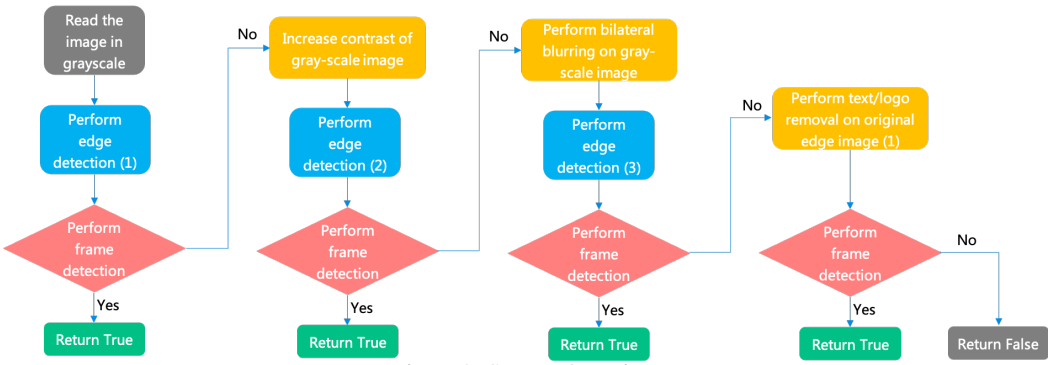


Figure 3: System Overview

and compared it with a deep learning model. However, their baseline algorithm is only capable of detecting frames instead of localizing them, let alone dealing with some more complex frames that involve texts or logos.

Text and Logo Detection and Removal

There have been a number of different approaches to detect and remove undesired texts or logos in images in past years. As for detecting texts in images, the ideas in [3]-[10] that combines both edge detection and connected component analysis on gray-scale images for detecting texts are similar to ours. Although they utilize various edge detection methods and different types of connected component analysis techniques, these algorithms all try to analyse properties of connected edge components like their geometric arrangements as well as height-width ratios to determine if they belong to certain text components. However, most of these methods do not include the process of removing the undesired texts and are targeted towards texts solely without considering logos that are almost always present, as least in online fashion images.

Our Method

In this section, we will describe our algorithm for frame component detection and localization in online fashion images. Figure 3 shows a system overview of our algorithm. With a sequential structure, we are able to concatenate several weak classifiers together to achieve a high performance. In our method, we have a total of four weak classifiers, where the first three are similar with slightly different but effective pre-processing methods. The fourth classifier involves a texts and logos removal process that targets those fashion images with superimposed texts or/and logos. Each input image will be first converted to a gray-scale image before being fed the classifiers.

Here are more details on our four weak classifiers. The first classifier simply consists of edge detection and frame detection. The second classifier differs slightly from the first one by supplementing a step of contrast increasing of the gray-scale image before edge detection. The third classifier differs from both the first and second classifiers by performing bilateral filtering [11] on the gray-scale image before edge detection. Bilateral filtering is a very effective blurring filter that can preserve edges at the same time. Therefore, it can help remove noise edges in the frames that result from textures or repeating patterns, while preserving important frame edges. The last weak classifier targets those images that contain interfering texts or logos in the frame regions. A text and logo removing procedure will first be performed on the edge image obtained in the first classifier to remove all connected text or logo edges, before detecting frames for the last time. At the end of each classifier, if any frame components are detected, the program will terminate and return the detected frame regions. Otherwise, it will proceed to the next classifier or return false at the last classifier, indicating that the input image does not contain any frames.

In the following subsections, key building blocks for our method will be described and explained in detail.

Edge Detection

In our method, a four-neighbour algorithm is adopted to obtain the gradient image G from the original gray-scale image X, followed by a fixed threshold classification to determine edge pixel. Equation 1 below shows how we calculate the gradient map:

$$G[r,c] = \tfrac{1}{4}|X[r-1,c]+X[r+1,c] \\ +X[r,c-1]+X[r,c+1]-4X[r,c]| \quad (1)$$

Then we use a fixed threshold $T = 20$ to obtain the edge map E based on the gradient image, following Equation 2.

$$E[r,c] = \begin{cases} 1, & G[r,c] \geq T \\ 0, & G[r,c] < T \end{cases} \quad (2)$$

Contrast Increment

This procedure is designed to detect frame boundaries that are too ambiguous to be easily discernible. In other words, the difference between neighboring pixels on the frame boundary is too small for our edge detection algorithm to distinguish. Therefore, in order to successfully detect these ambiguous edges, we apply a depolarizing filter to the original gray-scale image X to obtain an enhanced gray-scale image Y. The relationship between the original pixel value at a location (r,c), represented by $X[r,c]$, and the depolarized pixel value $Y[r,c]$ at the same location is shown in Figure 4. This process is defined by Equation 3.

$$Y[r,c] = \begin{cases} 2X[r,c], & X[r,c] \in [0,50] \\ 255-2(255-X[r,c]), & X[r,c] \in [205,255] \end{cases} \quad (3)$$

Bilateral Filtering

In some images, the frame regions may contain some textures or decorative patterns that look very much like non-frame regions in the edge image. Therefore, we adopt the edge-preserving

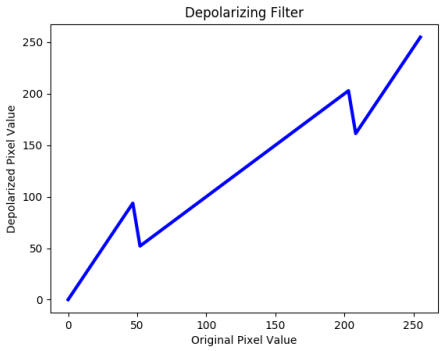


Figure 4: Depolarizing Filter

bilateral filter in our pipeline to try to detect this kind of frame. As shown in Figure 3, an edge detection process the same as in the previous classifiers is performed after applying the bilateral filter.

Text and Logo Removal

To effectively detect and remove edges corresponding to texts and logos in our extracted edge image, we propose a text and logo removal algorithm that is based on connected components, as part of our frame detection algorithm. To be specific, the algorithm aims to find all connected components that satisfy certain preset conditions in the edge image and then use multiple black rectangular boxes to cover these connected components. Connected components are searched over the entire edge image starting form the first pixel in the left upper corner in raster order. All components are connected based on a 4-neighbor relationship. Then, for each connected component, we will find its minimum and maximum row index, as well as its minimum and maximum column index among all pixels in the component, which will give us a rectangular region. By conditioning the height and width of this rectangular region, as well as the length of the connected component, we can decide whether this region belongs to a potential text or logo region. Finally, we remove the potential text or logo region by setting all pixel values in the region to zero. The conditions for a connected component to be recognized as text or logo region can be represented by the set of inequalities in Eq. 4. In these inequalities, R and C represent the set of row and column indices of all the pixels in a connected component, and H and W represent the height and width of the image in the unit of pixel. $\sum_{i=0}^{N} x_i$ represents the total number of pixels in a connected component.

$$\begin{aligned} 5 &\leq max(R) - min(R) \leq \frac{H}{2} \\ 5 &\leq max(C) - min(C) \leq \frac{W}{2} \\ \textstyle\sum_{i=0}^{N} x_i &\leq \frac{HW}{16} \end{aligned} \tag{4}$$

Fine-tuned ResNet-18

For purpose of comparison, we also fine-tuned a ResNet-18 [12] to detect frames as a binary classification problem. Since there are no publicly available datasets for frame detection, we need to build our own training dataset. To speed up the process, we first downloaded approximately 30000 fashion images randomly from Poshmark.com using a web image crawler. Then, we utilized our own frame detection algorithm that already has a good accuracy to filter out the images with frames, where we got about 8550 images. After that, we manually reviewed the images and kept only 8000 images that contains frames for sure. Then, we randomly selected the other 8000 images from the rest of approximately 20000 images that were not detected by our algorithm and formed the set of images that do not contain frames. As a result, we ended up with a training dataset containing 8000 images with frames and 8000 images without frame.

As for the training setup, we used stochastic gradient descent as the optimizer and the cross entropy loss function. We set the learning rate to be 0.01, and adopted a decaying learning rate of ratio 0.1 every 20 epochs. The training batch size was set to 16 images with each image resized to 224×224 pixels. The neural network was trained until convergence on a NVIDIA GeForce GTX 1080ti.

Experimental Results

To evaluate the performance of our proposed method and our fine-tuned ResNet-18, we gathered 1000 testing images that are completely differently from our training set from Poshmark.com. Among the 1000 testing images, there are 400 images containing frames and 600 not containing frames. All the testing images are real images uploaded by users of Poshmark. Some examples of the testing images are shown in Figure 5. The accuracy, true-positive rate, and true-negative rate are shown in Table 1.

(a) (b) (c)

(d) (e) (f)

Figure 5: Examples of testing images: (a)-(c):Images with frames; (d)-(e):Images without frames.

Table 1: Accuracy, True-Positive Rate, and True-Negative rate for Our Method and ResNet-18

Method	Accuracy	TP Rate	TN Rate
Our Method	93.6%	90.0%	96.0%
ResNet-18	94.1%	91.75%	95.83%

Both our proposed method and the fine-tuned ResNet-18 have very impressive accuracy. Although the overall accuracy of our proposed method is slightly lower than the fine-tuned ResNet-18, our proposed method achieves a higher true-negative rate at 96.0%, which means that our method is more cautious at detecting frames in images.

Conclusion

In this paper, we proposed a method consisting of a sequence of processing units and weak classifiers to detect frames in online fashion images. We also tried fine-tuning a Convolutional Neural Network (ResNet-18) as a comparison to our proposed method. Our experiments demonstrated the effectiveness of our proposed method and have shown that our proposed method achieved performance comparable to that of a fine-tuned ResNet-18 while having a slight advantage in the accuracy over Convolutional Neural Network in classifying images without frames.

Acknowledgments

We would like to thank all colleagues in our team who have provided assistance.

References

[1] C. Lu, J. Allebach, J. Wagner, B. Pitta, D. Larson, Y. Guo. Online image classification under monotonic decision boundary constraint. In Proc. of SPIE, Vol. 9395, pp. 9395-9395-8. (2015).

[2] allegro.tech, Deep learning for frame detection in product images, Dec. 19, 2016. [Online]. Available: http://https://allegro.tech/2016/12/deep-learning-for-frame-detection.html. [Accessed: Feb. 12, 2019].

[3] J. Gllavata, R. Ewerth, and B. Freisleben A robust algorithm for text detection in images, Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis. (2003).

[4] L. Agnihotri and N. Dimitrova. Text Detection for Video Analysis. In Proc. of the International Conference on Multimedia Computing and Systems, Florence, Italy, pp. 109-113. (1999).

[5] M. Cai, J. Song and M. R. Lyu. A New Approach for Video Text Detection. In Proc. of International Conference On Image Processing, Rochester, New York, USA, pp. 117-120. (2002).

[6] C. Garcia and X. Apostolidis. Text Detection and Segmentation in Complex Color Images. In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP2000), Istanbul, Vol. 4, pp. 2326-2330. (2000).

[7] A. K. Jain and B. Yu. Automatic Text Location in Images and Video Frames. In Proc. of International Conference of Pattern Recognition (ICPR), Brisbane, pp. 1497-1499. (1998).

[8] P. K. Kim. Automatic Text Location in Complex Color Images Using Local Color Quantization. IEEE TENCON, Vol. 1, pp. 629-632. (1999).

[9] R. Lienhart and W. Effelsberg. Automatic Text Segmentation and Text Recognition for Video Indexing. Multimedia System, Vol. 8, pp. 69-81. (2000).

[10] L. Hu, Z. Li, G. Gowala, S. Sundaram, P. Lee and J. Allebach. Non-native Content Detection and Localization for Online Fashion Images. In Proc. of Imaging and Multimedia Analytics in a Web and Mobile World 2019 (Part of IS&T Electronic Imaging 2019), Burlingame, CA. (2019).

[11] C. Tomasi and R. Manduchi. Bilateral Filtering for Gray and Color Images. In Proc. of the 1998 IEEE International Conference on Computer Vision, Bombay, India. (1998).

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

Author Biography

Litao Hu received his BS in Electronic Engineering from the Hong Kong University of Science and Technology (2017) and is currently a PhD candidate in Electrical and Computer Engineering of Purdue University. As a research assistant in the Electronic Imaging System Laboratory, his research interests include image processing, machine learning, and deep learning.

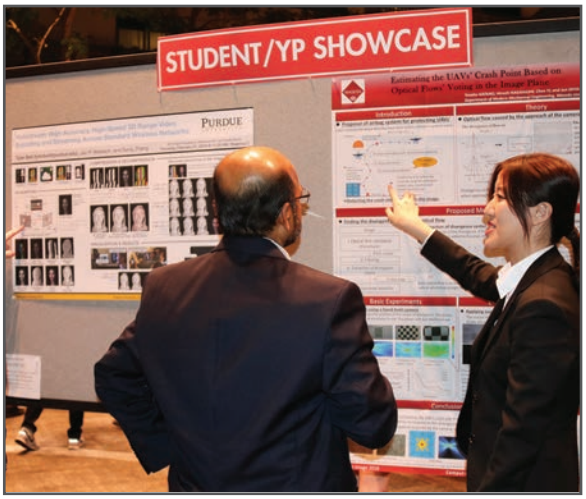
JOIN US AT THE NEXT EI!

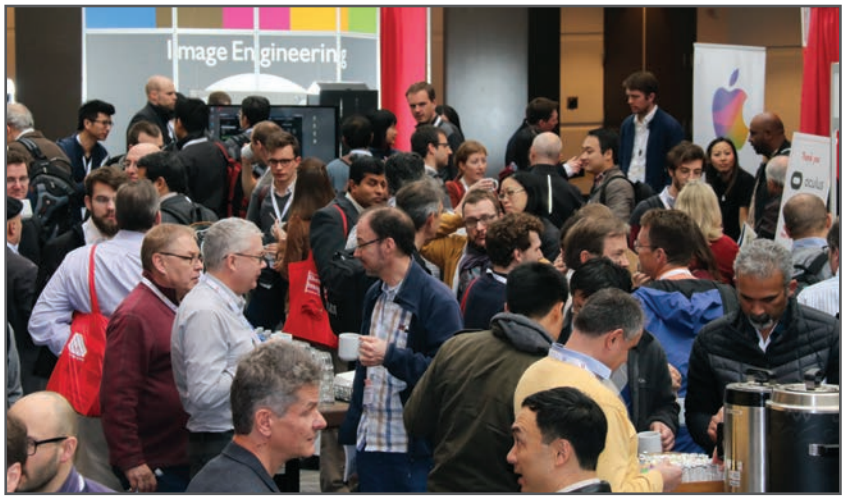
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org