Fashion Style in 128 Floats: Joint Ranking and Classification using Weak Data for Feature Extraction — SUPPLEMENTAL MATERIAL —

Edgar Simo-Serra
Waseda University
esimo@aoni.waseda.jp

Hiroshi Ishikawa
Waseda University
hfs@waseda.jp

Abstract

We present additional results that accompany the main paper submission. We provide additional visualizations of the learning process and the obtained descriptors. Details that were omitted from the submission due to space constraints are also included.

1. Dataset Cleaning Annotation Criterion

In order to minimize the noise of the Fashion144k [4], 6,000 images were annotated as suitable or not suitable for full body fashion style feature learning with the following criterion:

- Only one individual per image.
- Individual is fully visible (no crops).
- Individual is at least 40% of the image height in height.
- Individual is clearly identifiable and clothes is recognizable.

As this is a binary labelling task, a single image can be processed in a bit under a second by a user. This allows the annotation of the 6,000 images in less than two hours by a single person, which is sufficient to obtain 94.23% accuracy.

We follow the standard procedure for fine-tuning the VGG 16 layer [5], which consists of: removing the last layer, adding a new one with only two outputs with random initialization, and finally training with a much higher learning rate on the last layer. We show examples of positive predictions in Fig. 1 and examples of negative predictions in Fig. 2. We can see that in general it performs a very good task and thus it can be used automatically filter large amounts of data very cheaply. We use this procedure as the data obtained from

2. Siamese Architecture

A Siamese network [1] learns to recognize similar or dissimilar inputs. Our implementation takes two input images I_1 and I_2 and processes each with the feature extraction network to obtain two outputs $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. Afterwards the L_2 norm is computed on both outputs and a loss is applied that encourages similar inputs to have a small distance and dissimilar inputs to have a large distance. In particular we consider the Hinge embedding loss [3]:

$$l_S(\boldsymbol{f}_1, \boldsymbol{f}_2, s) = \begin{cases} \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_2, & \text{if} \quad s = 1 \\ \max(0, m - \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_2), & \text{else} \end{cases},$$

where $s = 1$ indicates the outputs should be similar, and m is a margin hyperparameter.

We perform learning in the same way as for the ranking only loss. Given the noisy labels $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ of image I_1 and I_2 respectively, we consider similar images $s = 1$ to be those where $r(\boldsymbol{y}_1, \boldsymbol{y}_2) > \tau_s$ and dissimilar images $s = 0$ to be those

Figure 1: Example results of clean images obtained from the dataset by filtering based on the prediction of a finetuned deep convolutional neural network. We can see in all cases there is a person that is the center of attention in the image.

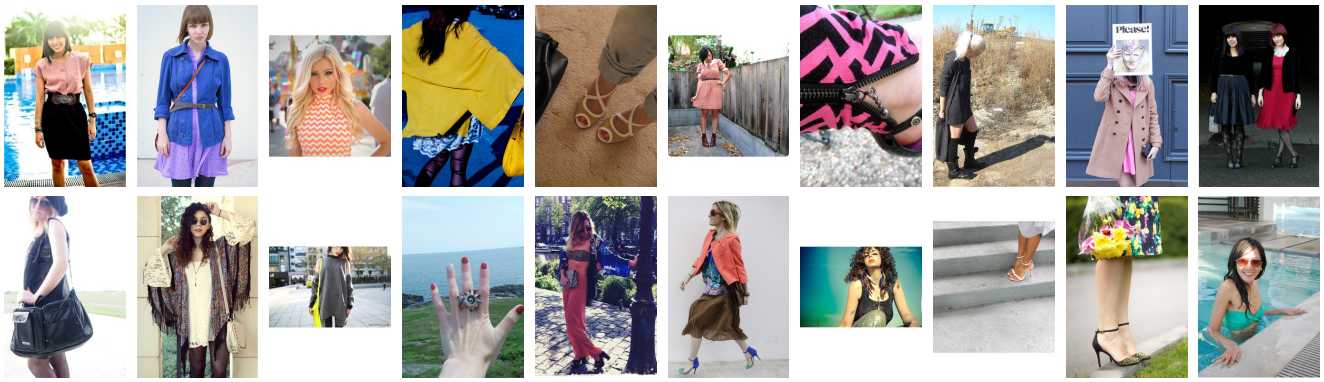

Figure 2: Example on non-clean images as predicted by our finetuned deep convolutional neural network. We can see that while there are a few mistakes, in general it does a very good job of getting rid of images that are not fit for learning such as extreme cropping and multiple people in the image.

Table 1: Confusion matrix for our joint classification and ranking model for 100 random splits with a 9:1 train to test ratio using the top $\delta = 0.5$ images from each style.

	Hipster	Goth	Preppy	Pinup	Bohemian
Hipster	1061	263	368	10	188
Goth	115	1906	96	22	52
Preppy	247	125	1588	57	144
Pinup	5	36	71	735	115
Bohemian	135	86	117	42	1916

where $r(\boldsymbol{y}_1, \boldsymbol{y}_2) < \tau_d$. We use the same values as in the ranking case, *i.e.*, $\tau_s = 0.75$ and $\tau_d = 0.01$. We use a margin of $m = 5$ for the Hinge embedding loss. For learning we randomly sample an equal number of positive and negative samples for each batch, and we initialize the weights with the best feature extraction network trained only on classification.

3. Hipster Wars Confusion Matrix

We additionally display the confusion matrix for our best performing approach (Ours Joint in the original paper) on the Hipster Wars [2] dataset in Table 1 for the 100 random splits with 9:1 train to test ratio for the top $\delta = 0.5$ images for each class. The model performs very well and most of the mistakes are between more easily confused classes such as Hipster, Preppy and Bohemian.

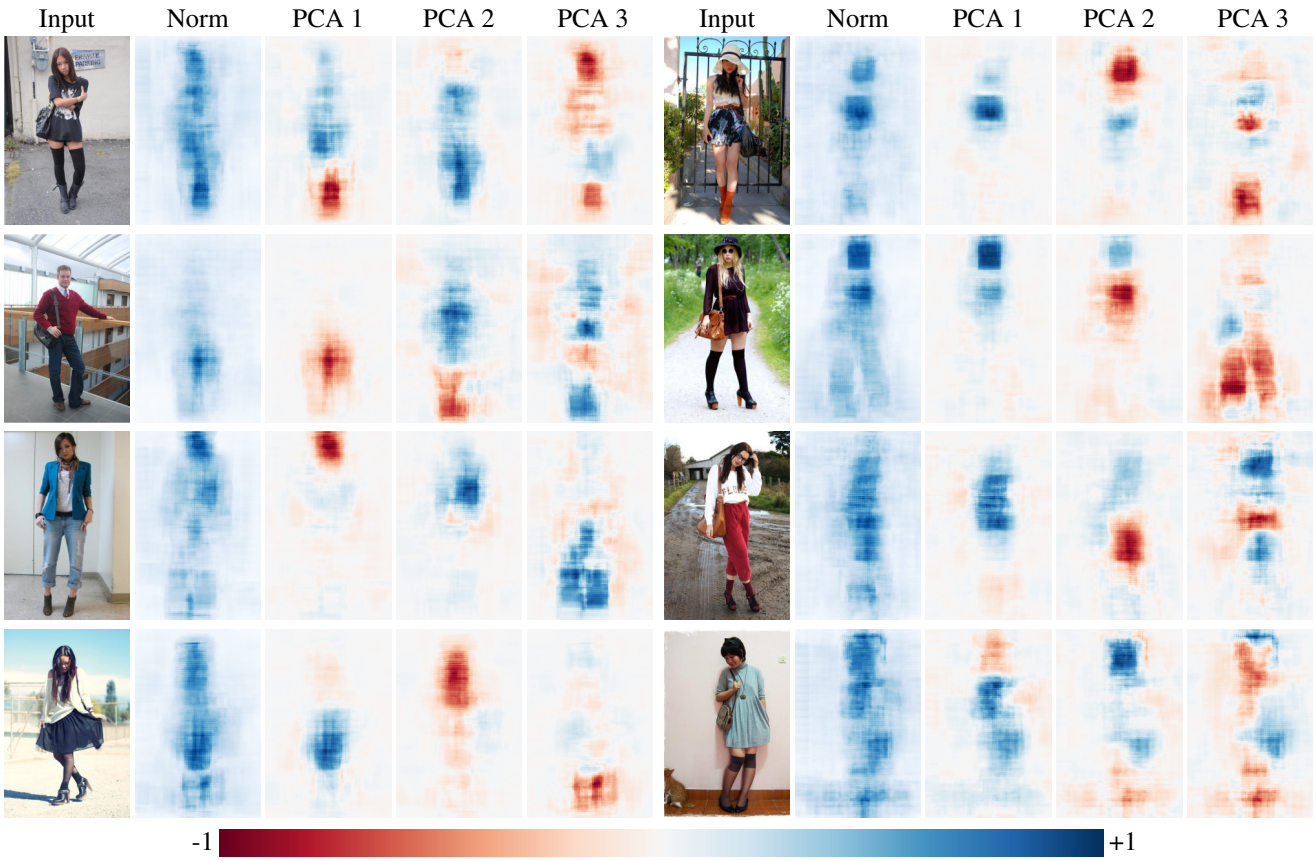

Figure 3: We analyze the relationship between the image and the style descriptor by moving an occluding box around the image and display the change in the norm of the descriptor and the change of the first three components on PCA basis computed on all the vectors extracted on the image. The positive and negative values are encoded in blue and red respectively. The norm is normalized so that the minimum value is white and the maximum value is blue, while the PCA representations are normalized by dividing by the maximum absolute value. Large descriptor changes correspond to the location of the individual and the different PCA modes refer to different locations of the body such as face, upper body, legs or shoes. Figure best viewed in color.

4. Visualizing the Style Descriptor

Additional visualizations of the style descriptor are shown in Fig. 3. We can see how the norm of the descriptor varies mostly with the body of the individuals, focusing on different aspects of their fashion. In general each PCA mode corresponds to a different aspect. The first mode (PCA 1) focuses on the dominant style part of the image, while later modes begin to mix more parts of the image and become noisier.

4.1. Style Space

A higher resolution visualization of the style space can be seen in Fig. 4. The style space is constructed by using the t-SNE [6] algorithm on the Euclidean distances between the descriptors computed from the different images as our ranking loss is directly optimizing for Euclidean distance.

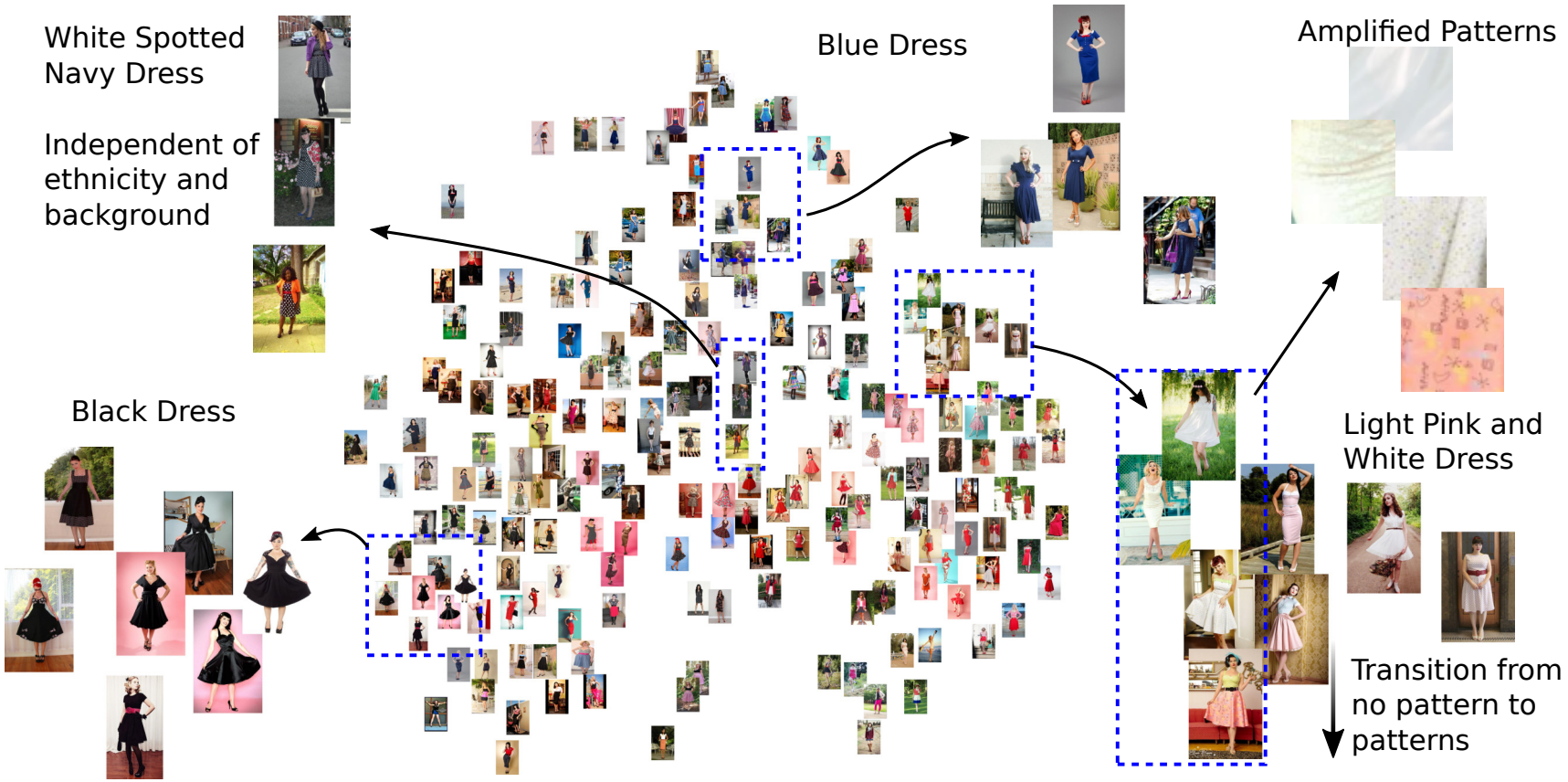

Figure 4: Visualization of the fashion style space of the Pinup class from the Hipster Wars [2] dataset using t-SNE [6]. Note the robustness to background and subjects. Figure best viewed in color.

References

[1] J. Bromley, I. Guyon, Y. Lecun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *NIPS*, 1994. 1

[2] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 2, 4

[3] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *ICML*, 2009. 1

[4] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. In *CVPR*, 2015. 1

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[6] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, Nov 2008. 3, 4