Mix and Match: Joint Model for Clothing and Attribute Recognition

Kota Yamaguchi[1]
http://vision.is.tohoku.ac.jp/~kyamagu

Takayuki Okatani[1]
okatani@vision.is.tohoku.ac.jp

Kyoko Sudo[2]
sudo.kyoko@lab.ntt.co.jp

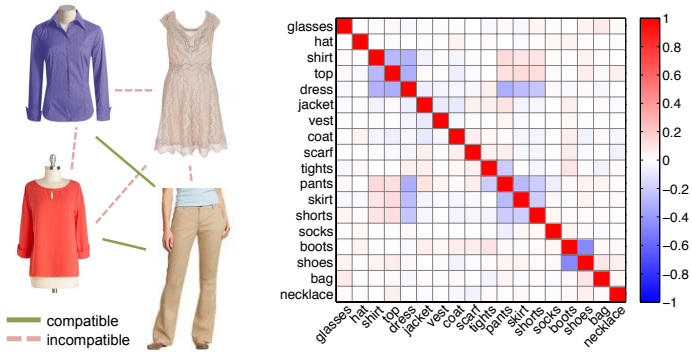
Kazuhiko Murasaki[2]
murasaki.kazuhiko@lab.ntt.co.jp

Yukinobu Taniguchi[3]
ytaniguti@ms.kagu.tus.ac.jp

[1] Tohoku University
Sendai, Japan

[2] NTT
Yokosuka, Japan

[3] Tokyo University of Science
Tokyo, Japan



(a) Matching clothing pairs (b) Pearson correlation between clothing items

Figure 1: We consider the compatibility of items and attributes in detection. In b), notice the exclusive blocks, e.g., *shirt*, *top*, and *dress*.

Abstract This paper studies clothing and attribute recognition in the fashion domain. Specifically, in this paper, we turn our attention to the compatibility of clothing items and attributes (Fig 1). For example, people do not wear a skirt and a dress at the same time, yet a jacket and a shirt are a preferred combination. We consider such inter-object or inter-attribute compatibility and formulate a Conditional Random Field (CRF) that seeks the most probable combination in the given picture. The model takes into account the location-specific appearance with respect to a human body and the semantic correlation between clothing items and attributes, which we learn using the max-margin framework. Fig 2 illustrates our pipeline. We evaluate our model using two datasets that resemble realistic application scenarios: on-line social networks and shopping sites. The empirical evaluation indicates that our model effectively improves the recognition performance over various baselines including the state-of-the-art feature designed exclusively for clothing recognition. The results also suggest that our model generalizes well to different fashion-related applications.

Joint detection Let us denote a set of labels by $Y \equiv \{y_i\},\ y_i \in \{0,1\}$, where i is one of the clothing items or attributes, such as *shirt* or *skirt:plain*. Given a feature $X \equiv \{\mathbf{x}_i\}$, we define our joint probability distribution over labels by a log-linear model:

$$\ln P(Y|X) \equiv \sum_i w_i \phi(\mathbf{x}_i, y_i) + \sum_{i,j \in V} w_{i,j} \psi(y_i, y_j) - \ln Z, \tag{1}$$

We use logistic regression of each label for the unary term, expressed by:

$$\phi(\mathbf{x}_i, y_i) \equiv \ln p(y_i|\mathbf{x}_i), \tag{2}$$

$$p(y_i = 1|\mathbf{x}_i) \equiv \sigma(\mathbf{a}_i^{\mathrm{T}} \mathbf{x}_i + b_i), \tag{3}$$

where $\mathbf{a}_i$ and b_i are the regression parameters for each item. For the binary term, we use the normalized Pearson correlation:

$$\psi(y_i, y_j) \equiv \begin{cases} \ln \frac{1}{2}\left(1 + c_{i,j}\right), & \text{if } y_i = y_j \\ \ln \frac{1}{2}\left(1 - c_{i,j}\right), & \text{otherwise} \end{cases} \tag{4}$$

where $c_{i,j}$ is the Pearson correlation between item i and j in the training examples.

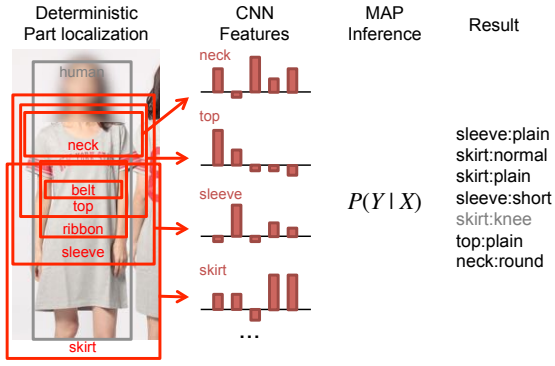
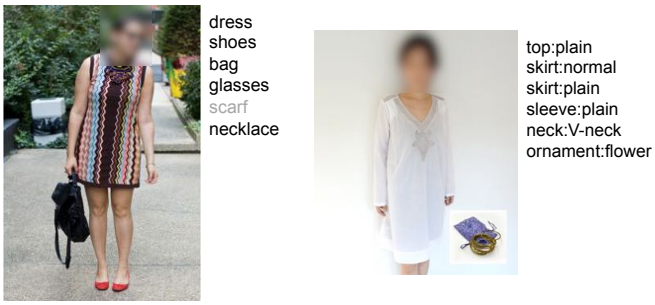


Figure 2: Our joint detection pipeline.



(a) Clothing detection (b) Attribute recognition

Figure 3: Results in a) clothing detection and b) attribute recognition.

Deterministically localized feature To reduce the dependency to very accurate pose estimation, we apply a simple yet effective approach for part localization. Given a bounding-box around the full human body, we extract an image patch from a specific location relative to the body bounding-box. Fig 2 illustrates an example of part bounding boxes for attribute detection. This deterministic approach only takes constant time yet produces surprisingly good results for fashion problems.

Once the region of interest is determined, we extract the CNN feature (`fc7` of AlexNet, 4096 dimensions) learned from ImageNet [1], and use as a feature-input to the logistic regression in Eq 2.

Experimental results We used two datasets: a) Chictopia dataset that considers automatic clothing tagging in fashion blogs, and b) Dress dataset that considers attribute recognition in e-commerce. Through the comparison between various baselines including the state-of-the-art image feature [2], we confirmed that the proposed joint model successfully improves the precision of prediction in both two scenarios. Fig 3 shows a few qualitative results.

[1] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014.

[2] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Retrieving similar styles to parse clothing. *TPAMI*, 2014.