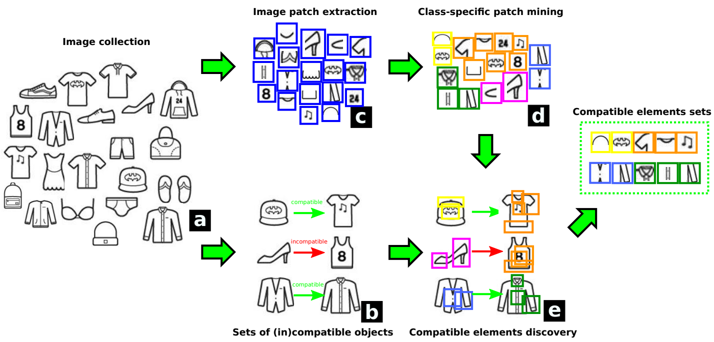
Estimating Inter-Class Visual Compatibility through Mid-level Elements

José Oramas M.
http://homes.esat.kuleuven.be/~joramas/
Tinne Tuytelaars
http://homes.esat.kuleuven.be/~tuytelaa/

KU Leuven, ESAT-PSI, iMinds,
Leuven, Belgium

1 Introduction

In this work we focus on measuring the visual compatibility of objects depicted in different images. Inspired by the success of representations based of mid-level elements [2, 3, 5, 8, 9], we propose a hierarchical method to learn visual representations in a data-driven fashion. At the base-level we discover a set of informative class-specific elements by mining activations of Convolutional Neural Networks (CNN). These base-level elements are effective at describing images of the classes of interest. At the top-level we exploit co-occurrences of base-level elements between images of compatible objects. This produces a set of "rules" which "explain" the compatibility of such objects. The main contribution of this work is a hierarchical method that not only measures the compatibility between objects depicted in images but also provides an insight on the features that drive the compatibility.



2 Proposed method

Our method follows a three-step approach as depicted in the figure above. Given a set of object images (a) accompanied with object-class labels, and links between them describing the (in)compatibility between them (b), we proceed as follows. First, we uniformly extract image patches for each image in the collection and compute CNN activations from each patch (c). Second, we perform mining on the CNN activations from images corresponding to each object class (d), producing a set of activation patterns (base-level elements) that describe each of the classes of interest. Third, given a set of image pairs, the base-level elements are detected in each image and pattern sets of compatible base-level elements are discovered (e) via association-rule mining [1].

2.1 Modeling Object Appearance via Base-level Elements

Base-level elements are extracted with the goal of describing different visual characteristics of instances of class c. Given an image of an object of class c, we uniformly sample a set of regions $R = \{r_1, r_2, \ldots, r_m\}$ and for each region r_i we compute CNN activations as an initial feature. Following this, we construct class-specific transaction-item matrices where each transaction is defined by an image region r_i^c from class c and the items are the CNN activations of such regions. Each matrix is binarized by selecting the top-k activations per transaction. This matrix is mined producing a set of patterns. Then, we remove redundant patterns that fire on the same set of items (CNN activation indices), thus, producing a reduced set of patterns P'^c. The next step consists of extracting a set of visual elements $V'^c \in P'^c$. The visual elements $v_i'^c \in V'^c$ cover all the image patches that contain the pattern $p_i'^c \in P'^c$. For each base-level element $v_i'^c$, we train an LDA classifier $\theta_i'^c$ using all the image patches covered by $v_i'^c$.

2.2 Modeling Visual Compatibility via Top-level Elements

Given a pair of images depicting compatible objects, the goal of top-level elements is to represent the set of object characteristics (described via base-level elements) that define such compatibility. To this end, given a set of image pairs with labels stating the compatibility of the depicted objects, first, we compute activations of base-level elements using the classifiers $\theta_i'^c$. Then, a new set of transaction-item matrices (one for each class combinations (c_i, c_j)) are defined where the transactions are defined by the image pairs and the items by the activations of base-level elements occurring on them. As a third step, each matrix is mined via association-rule mining [1] producing a set of patterns $P''^{(c_i,c_j)}$ which are further reduced to obtain the top-level elements $V''^{(c_i,c_j)} \in P''^{(c_i,c_j)}$. Finally, for each top-level element $v''^{(c_i,c_j)}$ we train a LDA classifier $\theta''^{(c_i,c_j)}$ using all the image pairs covered by $v''^{(c_i,c_j)}$. Each of these classifiers measures the level to which a specific top-level element occurs on a pair of images.

2.3 Inference

During testing, given a pair of images to be evaluated, we extract the activations of both base and top level elements using the classifiers $\theta'^{(c_i)}$ and $\theta''^{(c_i,c_j)}$, respectively. Then, we measure the compatibility by taking the maximum response between top-level elements. Please refer to [7] for a detailed description of the proposed method.

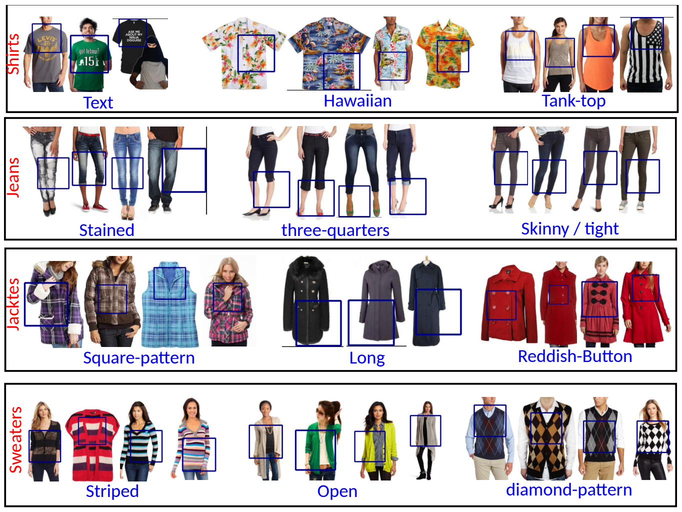


Figure 1: Some of the clusters defined by the base-level visual elements for the *shirts, jeans, coats & jackets*, and *sweaters* product classes. For each of the clusters we have added a text caption (in blue) of the style that they seem to encode. Furthermore, each image within each cluster is marked with the patch that encodes such style.

3 Evaluation

We conduct experiments on the clothes-related classes from the massive Amazon-based dataset collected in [6]. We focus our experiments on a subset covering ∼500K images using the compatibility annotations and image splits used in [10]. We compute CNN activations using *Caffe* [4] in combination with the *CaffeRef* model [4]. We consider a set of 4000 base/top elements at each level of the hierarchy. Total training time took below 12 hours. We compare w.r.t. the method proposed in [10] which also addresses the problem of measuring visual compatibility. The method from [10] operates at the image level by learning a space in which compatible objects are close via a siamese network architecture. Following the evaluation protocol from [10], we use as performance metric the area under the curve (AUC) defined by the False Positive Rate (FPR) vs. the True Positive Rate (TPR) on the compatibility estimation task. We show qualitative results in Fig. 1 and 3.

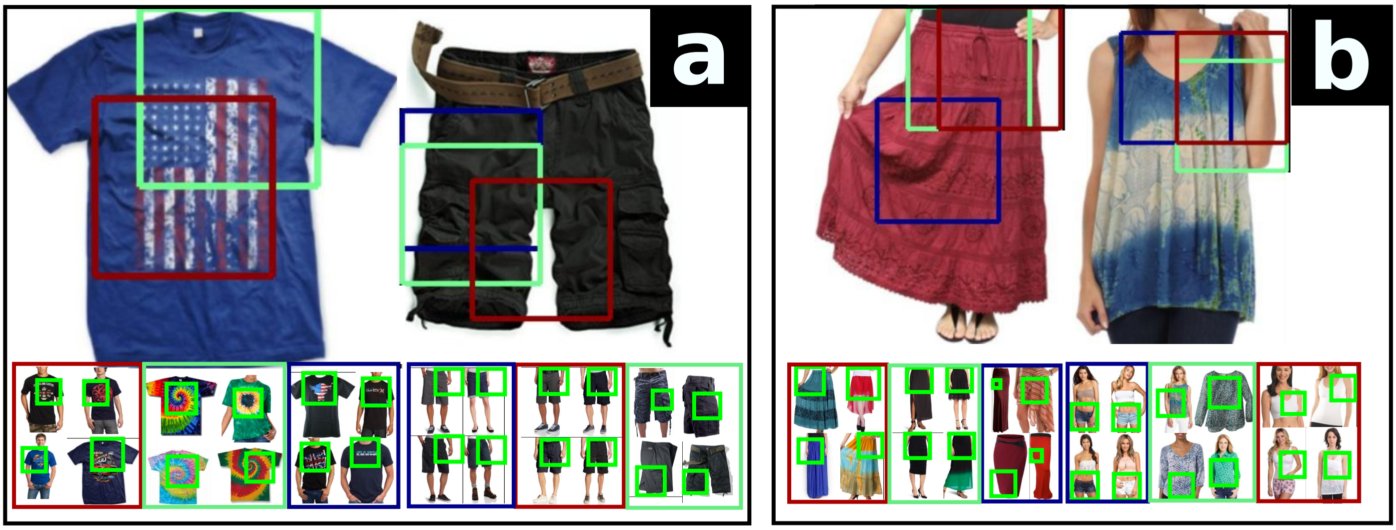


Figure 2: Accurately predicted compatible objects. For each object the regions of the top-3 base-level elements are indicated with their scores color-coded in jet scale (except for the t-shirt example where the 2^{nd} and 3^{rd} elements overlap). For each base-level element, we present a subset of random examples that compose it. Note how some base-level elements effectively describe some of the features that define the link between the objects.

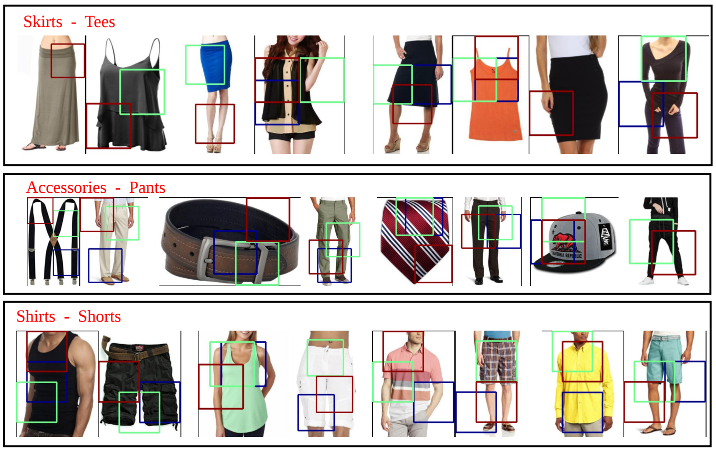


Figure 3: Some examples of accurately predicted compatible object pairs. For clarity, for each object the regions of only the top-3 base-level elements are indicated with their scores color-coded in jet scale.

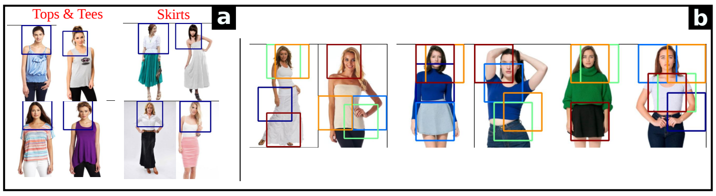


Figure 4: Mid-level elements encoding contextual information at the base (a) and top (b) levels, respectively. The left image shows the region (blue) of the base-level element. The right image shows comparisons between images with the top scoring base-level elements that describe the compatibility color-coded in jet scale. Note how the method is able to encode that *skirts* and some *tops* are female products (a). Moreover, detecting a female face can produce a strong link when comparing *skirts* and some *tops* products (b).

3.1 Discussion

On the quantitative side, the proposed method achieves subpar performance (0.66 AUC) when compared to the method from [10] (0.80 AUC). However, on the qualitative side, the method from [10] has a more reduced output in terms of "explaining" the reasons that make the compared objects compatible. This is mostly due to the fact that it operates in a black-box fashion with an output limited solely to a compatibility score. On the contrary, while still being able to estimate the compatibility up to some level, the proposed method not only effectively models different styles in which the objects of interest may occur (Fig. 1) , but it also provides an insight on the characteristics of the objects that define compatibility between them (see Fig. 2 & 3). Moreover, there is evidence (Fig. 4) suggesting that the method is able to exploit contextual information when modeling object classes and the compatibility between them. For example, the method is able to encode that *skirts* and some *tops* are female products (Fig.4.a). Likewise, it is able to encode that wearing 3/4 *jeans* leave the feet exposed (Fig.1, 2nd Row).

Despite its descriptive strength, there are several areas in which the proposed method can be improved. First, performing a sampling at different scales should bring improvement in two ways: i) better base-level element alignment, and ii) inspection of low-resolution base-level elements. This should provide more informative base-level elements to assist the discovery of top-level elements. Second, in its current state, the proposed method computes the compatibility score between images by taking the maximum response from the classifiers based on the top-level elements (Sec. 2.3). This removes the possibility of different top-level elements operating in a coordinated fashion. This is a clear weakness since some of these elements encode complementary properties, e.g. color, texture, shape. Future work will focus on improving the inference step by performing ensemble reasoning based on the top-level classifiers.

Acknowledgements: This work is supported by the PARIS project (IWT-SBO-Nr. 110067), the FWO project "Representations and algorithms for the captation, visualization and manipulation of moving 3D objects, subjects and scenes", iMinds and a NVIDIA Academic Hardware Grant.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, 1993.

[2] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

[3] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A Efros. What makes paris look like paris? *Communications of the ACM*, 58(12): 103–110, 2015.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *MM*, 2014.

[5] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Mining mid-level visual patterns with deep CNN activations. *CoRR*, abs/1506.06343, 2015.

[6] J. J. McAuley, R. Pandey, and J. Leskovec. Inferring networks of substitutable and complementary products. In *KDD*, 2015.

[7] J. Oramas M. and T. Tuytelaars. Modeling visual compatibility through hierarchical mid-level elements. *CoRR*, abs/1604.00036, 2016.

[8] K. Rematas, B. Fernando, F. Dellaert, and T. Tuytelaars. Dataset fingerprints: Exploring image collections through data mining. In *CVPR*, 2015.

[9] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[10] A. Veit, B. Kovacs, S. Bell, J. McAuely, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 2015.