Classification-Driven Active Contour for Dress Segmentation

Lixuan Yang[1,2], Helena Rodriguez[1], Michel Crucianu[2] and Marin Ferecatu[2]
[1]*Shopedia SAS, 55 rue La Boétie, 75008 Paris, France*
[2]*CEDRIC Lab, Conservatoire National des Arts et Métiers, 292 Rue Saint-Martin, 75003 Paris, France*
contact@shopedia.fr, firstname.lastname@cnam.fr

Keywords: Dress Extraction, Clothing Extraction, Segmentation, SVM Classification, Active Contour

Abstract: In this work we propose a a dedicated object extractor for dress segmentation in fashion images by combining local information with a prior learning. First, a person detector is applied to localize sites in the image that are likely to contain the object. Then, an intra-image two-stage learning process is developed to roughly separate foreground pixels from the background. Finally, the object is finely segmented by employing an active contour algorithm that takes into account the previous segmentation and injects specific knowledge about local curvature in the energy function. The method is validated on a database of manually segmented images. We show examples of both successful segmentation and difficult cases. We quantitatively analyze each component and compare with the well-known GrabCut foreground extraction method.

1 Introduction

Although the interest in developing dedicated search engines for fashion databases is several decades old (King and Lau, 1996), the field only started to develop with the recent massive proliferation of fashion web-stores and online retail shops. Indeed, more and more users expect online advertising to propose items that truly correspond to their expectations in terms of design, manufacturing and suitability. Localizing, extracting and tracking fashion items during web browsing allows the professionals to better understand the users' preferences and design web interfaces that make for a better web shopping experience.

In this work we propose a method to segment dresses in fashion images, which is the first step to solving a more difficult problem inspired by the need of professionals of online advertising and fashion media: present to the users relevant items from a database of clothes, based on the content of the web application they are consulting and its context of use. This goes far beyond the needs of a search engine: the user is not asked to interact with any search interface or formulate a query, but instead she is accompanied by automatic suggestions presenting in a non intrusive way a selection of products that are likely to interest her.

Many recent research efforts regarding fashion images deal with a quite different use case, that of meta search engines federating and comparing several online shops. These efforts focus on improving existing search engines to help users find products that match their preferences while preserving the "browsing" aspect (Datta et al., 2008). Online shops sometimes provide image tags for common visual attributes, such as color or pattern, but they usually form a proprietary, non heterogeneous and non standardized vocabulary, usually too small to characterize the visual diversity of desired clothing (Redi, 2013; Di et al., 2013). Moreover, in many cases users look for characteristics expressed by very subjective concepts and words, to describe a style, a given brand or a specific design. For this reason, much recent research work is focused in the development of detection, recognition and search of fashion items based in visual characteristics (Datta et al., 2008; Lew et al., 2006; Liu et al., 2007; Liu et al., 2012a).

Another approach models the target item based on attribute selection and high-level classification (Deselaers et al., 2008). In (Di et al., 2013), the authors train attribute classifiers on fine-grained clothing styles, formulating image retrieval as a classification problem, by ranking items that contains the same visual attributes as the input, which can be a list of words or an image. A similar idea is explored in (Hsu et al., 2011) where a set of features such as color, texture, SIFT features and object outlines are used to determine similarity scores between pairs of images. In (Chen et al., 2012), the authors propose

to extract low-level features in a pose-adaptive manner and combine complementary features for learning attribute classifiers by exploring mutual dependencies between the attributes by conditional random fields. To narrow the semantic gap between the low-level features of clothing and the high-level categories, (Liu et al., 2012a) propose to adopt middle-level clothing attributes (e.g., clothing category, color, pattern) as a bridge. More specifically, the clothing attributes are treated as latent variables in a latent Support Vector Machine (SVM) recommendation model. To perform on larger fine-grained clothing attributes, (Chen et al., 2015) proposed a novel double-path deep domain adaptation network for attribute prediction by modeling the data jointly from the unconstrained photos and images from large-scale online shopping stores.

A second approach consists in using part-based models to compensate the lack of pose estimation and model complex interactions in deformable objects (Felzenszwalb et al., 2010). To predict human occupations, (Song et al., 2011) use part-based models to characterize complex details and variable appearances of human clothing on the automatically aligned patches of human body parts, described by sparse coding and noise-tolerant capacities. A similar part-based model is proposed in (Nguyen et al., 2012) where image patches are described by a mixture of color and texture features. Parts are also used in (Liu et al., 2012b) to reduce the "feature gap" caused by human pose discrepancy, by using graph parsing technique to align human parts for cloth retrieval.

Another approach is based on segmentation and aggregation to select different cloth categories. In (Kalantidis et al., 2013), articulated pose estimation is followed by an over-segmentation of the relevant parts. Then, clustering by appearance creates a reference frame that facilitates rapid classification without requiring an actual training step. Body joints are incorporated in (Jammalamadaka et al., 2013) by estimating their prior distributions and then learning the cloth–joint co-occurrences of different cloth types as a part of a conditional random field framework to segment the image into different clothes. A similar idea is proposed in (Yamaguchi et al., 2015b), they formulate a CRF by inter-object or inter-attribute compatibility. (Simo-Serra et al., 2014) has exploited another way to formulate a CRF by taking into account the garment's priors, 2D body pose condition and limb segment. The framework in (Yamaguchi et al., 2015a) is based on bottom-up clothing parsing from semantic labels attached to each pixel. Local models of clothing items are learned on-the-fly from retrieved examples and parse mask predictions are transferred from these examples to the query image. Face detection is used in (Yang and Yu, 2011) to locate and track human faces in surveillance videos, then clothing is extracted by Voronoi partitioning to select seeds for region growing. For the video applications, (Liu et al., 2014) use Sift Flow and super-pixel matching to build correspondences across frames and exploit the cross-frame contexts to enhance human pose estimation. Also for the human parsing and pose estimation, (Dong et al., 2014) proposed an unified framework to formulate the problem jointly via a tailored And-Or graph.

(a) (b) (c)

Figure 1: We aim to produce a precise segmentation of the fashion items as in (b). Recent state of the art (Yamaguchi et al., 2015a) produces the result in right figure (c), which is insufficient to provide a precise description of the object (dress in this case).

Unlike the methods described above, our proposal aims to segment *precisely* the object of interest from the background (foreground separation, see Fig. 1(b)), a difficult problem without user interaction and without using an extensive training database. Indeed, to propose meaningful results in terms of high-level expectations (such as product style or design) we need to achieve a good description of the visual appearance, which is much better if the object is segmented to eliminate the effect of mixing with the background and to include the shape outline in the description.

As an example, we show in Fig. 1(c), the results obtained on the left image by using state-of-the art method from (Yamaguchi et al., 2015a). Even though this method is capable of multi-label segmentation, the result of the segmentation is inadequate for a fine description of the dress. Instead, we aim to obtain a finer description like the one in Fig. 1(b) (which, incidentally, is produced by the method described in this paper).

To achieve our goal, we combine a person detector with a two stage SVM classification to achieve a rough estimation of the cloth contour (separation of the object from the background). The result is then used to seed an active contour, fine-tuned by prior specific knowledge of the object structure.

The novelty of our method is twofold: first, we combine local information in the image with a learning prior to guide the segmentation (which allows to predict the contour in the presence of small occlusions, rather than just following local contours) and second, we inject specific knowledge about the object (local curvature) in the energy function that guides the convergence of the active contour, which helps disambiguate the object's contour in cluttered background. The procedure requires a training database for the person detector and for the foreground detection stage, together with the collection of prior knowledge regarding the object to be segmented. To keep the presentation uncluttered, in this paper we focus on dresses, a challenging class to segment because of the complexity of the object and the variety of the environment in real images. However, our procedure can be adapted to any fashion cloth item by following a similar development approach.

Evaluation tests performed on a database of 200 manually segmented images show very promising results. We provide examples of successful segmentation, analyze difficult cases, and also evaluate quantitatively each component. Because existing methods, to our knowledge, do not segment precisely fashion items at this moment, we compare our method to the GrabCut (Rother et al., 2004) foreground extraction method, which is well-known in our community, is frequently used as a baseline case in many research works and has an open source implementation [1].

The rest of the paper is organized as follows: In Sec. 2.1 we give an overview of our proposal, followed by a detailed presentation of each component: person detector in Sec. 2.2, SVM-based detection in Sec. 2.3 and active contour in Sec. 2.4. After the experimental validation is presented in Sec. 3, we conclude the paper in Sec. 4 by a discussion of the main points and perspectives for further extensions.

2 Our proposal

Detecting dresses in images is a difficult problem because the object is deformable, can be composed of a large variety of materials, textures and patterns, and shows great differences in style and design inside this class. Also, it can appear against very different and complex backgrounds.

2.1 Overview of the approach

Since we aim to find as precisely as possible the contour of the dress, a direct approach is deemed to fail because of the aforementioned difficulties. Instead, we adopt a three stage method, each step preparing the following one as follows:

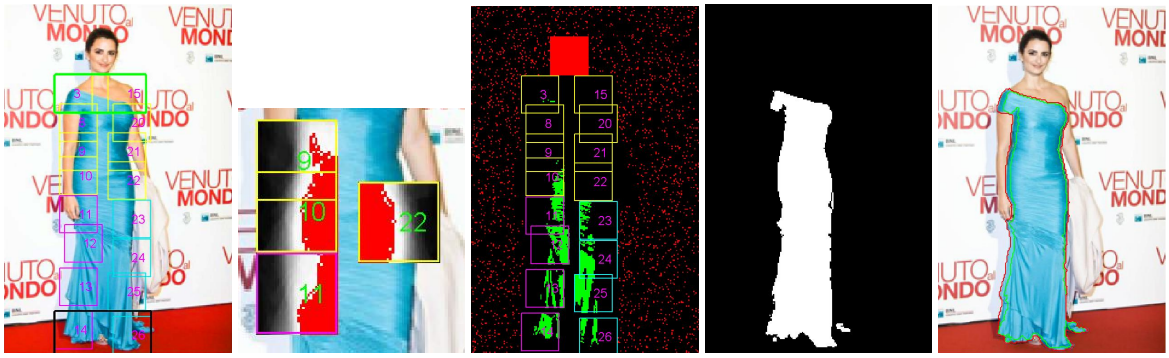
1. **Person detector.** We first train a person detector on a manually annotated database to find the regions of the image most likely to contain the contour of the object. We use the articulated human detection model with a flexible mixture of parts presented in (Yang and Ramanan, 2013), which works well for both person detection and pose estimation, and has been tested with success in several other fashion-related works (see Sec. 1). The output of the person detector is a set of parts (rectangular boxes) centered on the body joints and oriented correctly.
2. **Coarse foreground detection.** We employ the training data to estimate a probability map that each pixel inside a box belongs to the object. The map is used to seed a one-class SVM estimating the support of the distribution of positive examples (pixels that belong to the object). Then, a two-class SVM is trained to improve the (coarse) detection of pixels belonging to the object, taking as negative examples random background pixels.
3. **Active contour.** The result of the two-class SVM is used as input to a two step active contour procedure that produces the final segmentation. We include several specific terms in the energy function that guides the active contour: the first term uses the results of the learning stage to push the contour towards the SVM separation frontier (i.e., the contour of the object according to the SVM), the second term takes into account the local curvature weighted by the location on the object. This ensures a good balance between the local pixel behavior and the information injected by learning, producing good results in most situations.

2.2 Person detector

For clothing detection and recognition, applying a person detector is a reasonable starting point. As in many other studies (e.g. (Liu et al., 2012b; Kalantidis et al., 2013; Yamaguchi et al., 2015a)), we use the person detection with articulated pose estimation algorithm from (Yang and Ramanan, 2013), which has been extensively tested and proved to be very robust in many practical situations. It is based on a deformable model that defines the object as a combination of parts (Yang and Ramanan, 2013). The detection score is defined as the fit to parts minus the parts deformation cost. The mixture model not only encodes object structure but also captures spatial relations between part locations and co-occurrence relations between parts.

To train the person detector, we manually annotate

[1] http://opencv.org/



(a) (b) (c) (d) (e)

Figure 2: Different stages of our approach: (a) output of the person detector, (b) output of the one class SVM, (c) and (d) input and output of the two class SVM, (e) final result after first (green) and second (red) active contour steps

a database of 100 images of dresses. Each person in a training image is annotated with 14 joint points by marking the articulations and main body parts. The legs are usually covered by the long dresses, so the lower parts are placed on the edges of the dress rather than on the legs. This not only improves detection accuracy, but also hints to the location of the contours. Given an image I, the person detector provides a set of 26 body parts, each part being a square region resized to 40×40 pixels. Each part has a symmetric part with respect to the vertical axis and corresponds to a body part or articulation. In Fig. 2(a) we see the output of the person detector on an unannotated image. Note how boxes slightly overlap at each end most of the time. To reduce the search space, we take advantage of the fact that the dress contour is located inside the boxes and we close the outline by adding the green (upper) and black (bottom) boxes to make the connection between the left and right body part estimations. These new boxes are computed to fit the internal and external hull of the existing ones. All subsequent processing of each image is done only in the region outlined by the boxes.

Probability map. For each location (pixel) in every box, we compute the probability of occurrence of the dress. This probability map indicates for each box where the dress is more likely to be found. We use this map to harvest positive examples for the next stage (SVM classification, see Sec. 2.3). To compute the map, we manually segment the dress in all the training images and then, for each pixel position in each box b, we compute the value:

$$p_b(i,j) = \frac{\sum_{k=1}^{n} \delta(I_{kb}(i,j) \in \text{Dress})}{n}$$

where n is the number of training images and $\delta(I_{kb}(i,j) \in \text{Dress})$ is 1 iff pixel (i,j) from box b in image I_k belongs to the dress.

2.3 Coarse foreground detection

In the second stage, for each box we train a two-class SVM (Scholkopf and Smola, 2001) to separate foreground pixels (dress) from the background by using the prior information given by the probability map and the person detector. Each pixel is described by the RGB coordinates concatenated with other local characteristics as described in Sec. 3. In a new unseen image we only have the result of the person detector to start with (the 26 part boxes). Using directly the probability map computed earlier to find positive training examples by thresholding does not yield enough good examples to guarantee correct generalization. Instead, we use the 100 most probable pixels in the box (according to the probability map) to train a one-class SVM that computes the support of the positive examples in the input space. We observed experimentally that a number larger than 100 reduces the generalization ability of the resulting classifier. The main source of problems here is that pixel classification is prone to local instability in cluttered scenes. To counter this, we use the *context of each part* to inject confidence into the decision by allowing the neighboring parts to vote. Concretely, for each part, we build four one-class SVMs: one for the part itself, one for the symmetric part w.r.t. the vertical and two for the lower and upper neighboring parts (see Fig. 2(b)). A pixel is considered positive is all four one-class SVMs validate it.

Merely using the output of the previous one-class classification fails to isolate properly the dress because pixels from the background and from the dress may have similar descriptors. We thus take some of the background pixels as negative examples for a two class SVM. More precisely, the pixels predicted by the one-class classifier as belonging to the dress are considered as positive examples. We randomly take as negative examples an equal number of pix-

els from the background (outside the envelope of all the parts) to obtain a balanced training problem. We also include the head part in the negative examples. In Fig. 2(c) we illustrate the training set for the two-class SVM and in Fig. 2(d) we show the result of prediction. It can be seen that the cloud of positive predictions outlines the dress quite closely, meaning the learning problem is well posed.

2.4 Active Contour

The score of the two-class SVM on a pixel indicates the likelihood of the dress presence. To get the final dress segmentation, we use the active contour (AC) introduced in (Chan and Vese, 2001), a model that can segment objects whose boundaries are not necessarily well-supported by gradient information. The AC minimizes an energy that drives the evolution of the active contour towards the desired boundary. The energy of the contour is defined by its length, area and a fitting term:

$$F(c_1,c_2,C) = \mu * \text{Length}(C) + \nu * \text{Area}(\text{inside}(C)) + \quad (1)$$
$$+\lambda_1 \int_{inside(C)} |u_0(x,y) - c_1|^2 +$$
$$+\lambda_2 \int_{outside(C)} |u_0(x,y) - c_2|^2$$

where C is the current contour, c_1 and c_2 are the average pixel gray level values $u(x,y)$ inside and respectively outside the contour C. The curvature term is controlled by μ and the fitting terms by λ_1 and λ_2.

To achieve a faster convergence to the final result, a two-step procedure is employed. A first AC is initialized with the external envelope of the parts produced by the person detector and converges rapidly to an approximation of the desired boundary. It takes as input the binary image produced by the two-class SVM and requires a small μ allowing for strong curvatures. Then, a second AC is used to converge to the real contour. It is initialized with the contour produced by the previous step and has as input the gray level image. This step is using prior information about the object: the curvature in various areas is expected to be different. In some images the hand will cover the dress, we thus give a large value to μ in those parts to obtain a smoothed result. On the contrary, the shoulder, the lower part of the dress and the elbow may have strong curvatures, so we set a small value to μ in these parts to correctly follow the contour. For the other parts, we take a medium value for μ.

The mean values c_1 and c_2 are computed usually on the entire image. Because of the large variability of the background in real images, this values is meaningless in a local context in our case and we replace them by the average values calculated in a local window of size 40×40 pixels around each contour pixel.

To reinforce the role of the SVM-based classification on the position of the AC, we include a new term in the energy function (Eq. 1) that pushes the contour towards the SVM separation frontier:

$$F_{svm}(C) = \eta \int_{on(C)} |f_{svm}(x,y)|^2 \quad (2)$$

where f_{svm} is the two-class SVM decision function mapped between 0 and 1 by the logistic function $1/(1+e^{-|\cdot|})$.

3 Experimental results

At this time, we found no databases dedicated to evaluating cloth extraction/segmentation from natural images that could be used with our framework, i.e. which has enough number of dresses (our object of interest) to make training feasible. The closest we found is Fashionista (Yamaguchi, 2015; Yamaguchi et al., 2013) which is build to test accuracy of multi-label assignment to pixels, but their method (Paper Doll Parsing) trained on this database did not perform very well for extracting long dresses (see Fig. 1 and Fig. 3 for some examples).

(a) (b) (c)

Figure 3: Paper Doll (Yamaguchi et al., 2013) is inadequate for the precise segmentation needed by our use case: (a) original image, (b) our method, (c) Paper Doll.

Instead, we evaluate our method on a database of 200 manually segmented dress images, half of which are used for training and the other half for testing. This is enough for preliminary testing of the method, but of course a larger database is needed for full validation, also including other fashion objects. We plan to do this in an extension of the present work.

For the SVM prediction stage, we describe each pixel by its RGB values concatenated with the x and y derivatives. To correctly separate the dress from the background, especially when the two have similar colors, we further concatenate with the frequency

(a) (b) (c) (d) (e) (f)

Figure 4: Qualitative evaluation: (a, c, e) original images; (b, d, f) segmentation results.

(a) Original image (b) Our method (c) GrabCut

Figure 5: Visual comparison with GrabCut.

distribution of the power spectral density computed in a patch of size 8×8 pixels around the pixel. This is a well-known texture descriptor (Deselaers et al., 2008). Concerning the SVM, we used the LibSVM implementation (Chang and Lin, 2011) with $C = 100$ and the Gaussian kernel with scale $\gamma = 0.1$, parameters obtained by cross-validation on the training data.

In regard to the computational efficiency of our method, the time needed to extract the object contour from an image of size 800x600 pixels in of the order of 5 seconds on an average PC. This could likely be improved by a factor of 5 to 10 by parallel computation and code optimization. However, in our application scenario this is not needed, because the extraction is not real-time, so we did not pursue further this direction.

Quantitative evaluation: We compare the segmentation produced by our method to the one provided by a human. As performance measures we use the average rates of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) pixels, traditionally used for classifiers. To show the impact of each component, in Table 1 we compare the segmentation obtained by the original AC (Chan and Vese, 2001), the classification obtained by the SVM with and without the original AC and our method (SVM classification followed by the two step active contour with energy function enhanced by curvature compensation and SVM regularization terms). We see that both the active contour alone and the SVM alone produce few false positives, but a very high rate of FN, i.e. tend to miss-classify dress regions as background. The full method achieves a more adequate behavior in most cases: rather low error rates because of the curvature and learning compensation terms in the active contour stage.

Method	TP	TN	FP	FN
SVM	65.22	95.42	4.58	34.78
Original AC	79.41	93.64	6.36	20.59
Original AC + SVM	84.2	89.91	10.09	15.80
Full Method	87.06	90.3	9.7	12.94
GrabCut	93.72	52.29	47.71	6.28

Table 1: Evaluation of different configurations for our method and comparison with GrabCut.

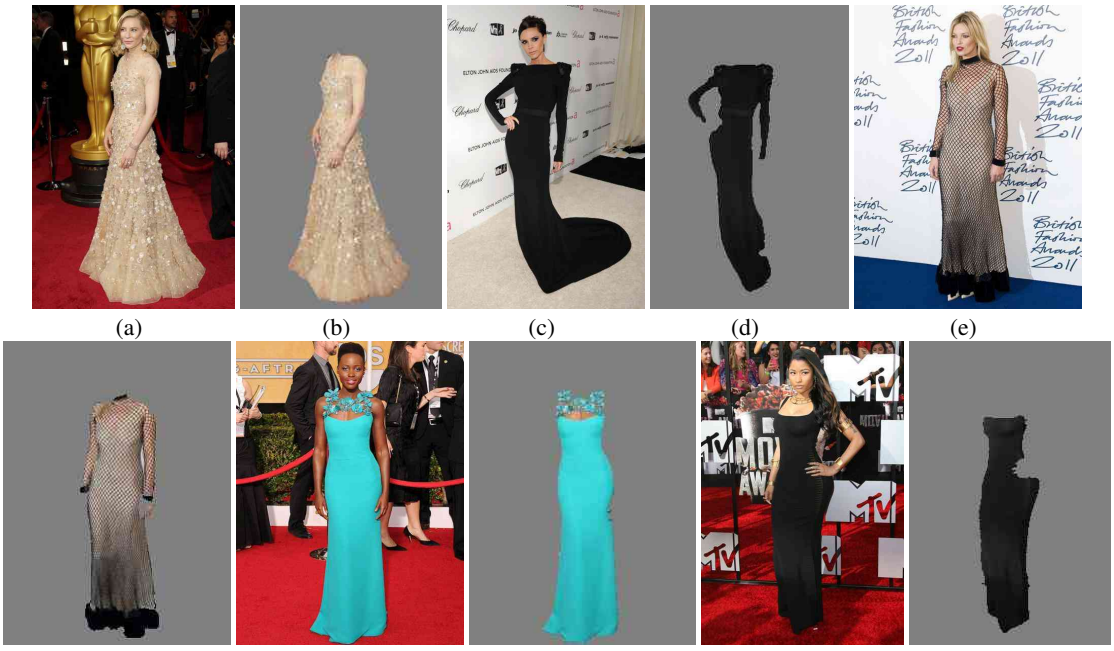
In a second set of experiments we compare our method with GrabCut (Rother et al., 2004), a well-known method for foreground extraction (see Table 1). To outline the object for GrabCut we used the external envelope of the parts identified by the person detector. GrabCut has a good rate of true positives (i.e. good classification of dress regions) but a too high FP rate, i.e. tendency to classify background as dress. This is likely due to the fact that GrabCut performs better for extracting objects on an uniform background, while out database contains many cases where the background is complex or dress and background are visually similar (see Fig 4).

We also evaluate the segmentation by means of the Jaccard (Intersection-Over-Union) score, score more frequently used in papers dealing with image segmentation:

$$S = \frac{\text{Surface}(Y \cap Y')}{\text{Surface}(Y \cup Y')}$$

Also for this measure, our method (79.7%) largely outperform GrabCut (64.86%).

Qualitative evaluation. In this part we illustrate the preceding conclusions with some examples taken from the test database. A first example of successful segmentation was already shown in Fig. 2(e). In Fig. 4 we present some other difficult examples of successful segmentation: (a,b) semi-transparent dress against



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 6: More results on different types of background.

skin color and (c,d) black dress against cluttered background with people dressed in black. In Fig. 4(e, f) we see a case of less successful segmentation: a red dress on a red carpet. Here, the pixel descriptors are not sufficiently discriminant to separate the foreground. More examples are presented in Fig. 6 to illustrate the behavior of the method with respect to different types of background.

In Fig. 5 we show a visual comparison with GrabCut: as hinted by the quantitative results, GrabCut includes many background pixels in the foreground. This occurs on all the database, explaining the high FP rate in Table 1.

4 Conclusion

We presented a novel method for dress segmentation that injects specific knowledge about the object into a three stage detection model combining learning and active contours. The inclusion of more training images, both for the person detector and for the probability map, together with the use of more sophisticated pixel descriptors should allow to further improve the results. In an extension of this work, we intend to evaluate and adapt the method for other types of clothing items and, by replacing the person detector with a deformable parts model (Felzenszwalb et al., 2010), to other fashion objects.

REFERENCES

Chan, T. F. and Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277. 5, 6

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. 6

Chen, H., Gallagher, A., and Girod, B. (2012). Describing clothing by semantic attributes. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 609–623. 1

Chen, Q., Huang, J., Feris, R., Brown, L. M., Dong, J., and Yan, S. (2015). Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*. 2

Datta, R., Joshi, D., Li, J., and Wang, J. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):1–60. 1

Deselaers, T., Keysers, D., and Ney, H. (2008). Features for image retrieval: An experimental comparison. *Information Retrieval*, 11(2):77–107. 1, 6

Di, W., Wah, C., Bhardwaj, A., Piramuthu, R., and Sundaresan, N. (2013). Style finder: Fine-grained clothing style detection and retrieval. In *Third IEEE International Workshop on Mobile Vision, CVPR*, pages 8–13. 1

Dong, J., Chen, Q., Shen, X., Yang, J., and Yan, S. (2014). Towards unified human parsing and pose estimation. In *CVPR*, pages 843–850. 2

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transac-*

tions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645. 2, 7

Hsu, E., Paz, C., and Shen, S. (2011). Clothing image retrieval for smarter shopping (stanford project). 1

Jammalamadaka, N., Minocha, A., Singh, D., and Jawahar, C. (2013). Parsing clothes in unrestricted images. In *Proceedings of the British Machine Vision Conference*. 2

Kalantidis, Y., Kennedy, L., and Li, L.-J. (2013). Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ICMR '13, pages 105–112. 2, 3

King, I. and Lau, T. K. (1996). A feature-based image retrieval database for the fashion, textile, and clothing industry in hong kong. In *International Symposium on Multi-Technology Information Processing (ISMIP'96), Hsin-Chu, Taiwan*. 1

Lew, M., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):19. 1

Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., and Yan, S. (2012a). Hi, magic closet, tell me what to wear! In *ACM Multimedia*, pages 619–628. ACM. 1, 2

Liu, S., Liang, X., Liu, l., Lu, K., Liang, L., and Yan, S. (2014). Fashion parsing with video context. In *ACM Multimedia*, pages 467–476. ACM. 2

Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., and Yan, S. (2012b). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337. 2, 3

Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282. 1

Nguyen, T. V., Liu, S., Ni, B., Tan, J., Rui, Y., and Yan, S. (2012). Sense beauty via face, dressing, and/or voice. In *ACM Multimedia*, pages 239–248. 2

Redi, M. (2013). *Novel methods for semantic and aesthetic multimedia retrieval*. PhD thesis, Universite de Nice, Sophia Antipolis. 1

Rother, C., Kolmogorov, V., and Blake, A. (2004). GrabCut – interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, pages 309–314. 3, 6

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press. 4

Simo-Serra, E., Fidler, S., Moreno-Noguer, F., and Urtasun, R. (2014). A high performance crf model for clothes parsing. In *ACCV*. 2

Song, Z., Wang, M., sheng Hua, X., and Yan, S. (2011). Predicting occupation via human clothing and contexts. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1084–1091, Washington, DC, USA. IEEE Computer Society. 2

Yamaguchi, K. (Last checked: September 2015). Fashionista image database, http://vision.is.tohoku.ac.jp/~kyamagu/research/clothing_parsing/. 5

Yamaguchi, K., Hadi, K., Luis, E., and Tamara, L. B. (2015a). Retrieving similar styles to parse clothing. *IEEE Transaction on PAMI*, 37:1028–1040. 2, 3

Yamaguchi, K., Kiapour, M. H., and Berg, T. L. (2013). Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 3519–3526, Washington, DC, USA. 5

Yamaguchi, K., Okatani, T., Sudo, K., Murasaki, K., and Taniguchi, Y. (2015b). Mix and match: Joint model for clothing and attribute recognition. In Xie, X., W. Jones, M., and K. L. Tam, G., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 51.1–51.12. BMVA Press. 2

Yang, M. and Yu, K. (2011). Real-time clothing recognition in surveillance videos. In *ICIP*, pages 2937–2940. IEEE. 2

Yang, Y. and Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transaction on PAMI*, 35:2878–2890. 3