Statistical and Deep Learning Algorithms for Annotating and Parsing Clothing Items in Fashion Photographs

Keiller Nogueira, Adriano Alonso Veloso, Jefersson A. dos Santos

Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Email: {keillernogueira, adrianov, jefersson}@dcc.ufmg.br

***Abstract*—Clothing identification has important roles in several areas. In this work, we present effective algorithms to automatically annotate and parse clothes from social media data. Clothing annotation tries to recognize each garment item that appears in a photo. Clothing parsing, in turn, locates and annotates each garment item in a photo. Both task pose interesting challenges for existing vision and recognition algorithms, such as distinguishing similar clothes or creating a pattern of a specific item. For the first task, two approaches, based on traditional algorithms, were proposed: (i) the pointwise one, and (ii) a multi-instance or pairwise approach. An evaluation show improvements of the proposed methods when compared to popular first choice algorithms that range from 20% to 30%. For the second task, a multi-scale convolutional network was proposed. At the end, a class is associated with each patch of the image. Experiments shows that the proposed method achieves promising results.**

***Keywords*-Machine Learning; Image Annotation; Image Parsing; Descriptor; Visual Dictionary; Neural Networks; Deep Learning;**

I. Introduction

Clothing parsing and annotation play important roles in human pose estimation [1], action recognition, person search [2], [3], surveillance [4], cloth retrieval [5] and have applications in fashion industry [1]. Considering the last one, applications with fashion images gained a lot of visibility with the increase of social networks and the faster spread of information, since these networks allow their members to express themselves in different ways, by creating and sharing content, making, for example, a new trend more successful or not. A particular way of expression being increasingly adopted is to post photos showing their latest looks and clothes. There are even specific networks for this, such as pose.com and chictopia.com. These social media channels carry a lot of information that, when analyzed, may help retailers and e-commerce systems to capture new trends helping to define new products and sales. To do so, it would be essential to find out the most popular clothes and in which segment they have been used more. Recommendation systems could also use this information to suggest new clothes based on searches already made or in the wardrobe of the users.

Although interesting, to reach suitable results for clothing applications it is necessary to extract all feasible information from the data, and this is only achieved with images entirely prepared, i.e., images fully annotated or segmented. However, only a very small percentage of images collected from social media have been associated with its clothing content [6], and manual methods are too expensive and maybe impracticable given the total amount of images. So, automatic algorithms appear as a very appealing alternative to reduce costs, but with difficult challenges to overcome. One challenge would be to differ similar types of garment items. For example, discerning a shirt from a coat is a very difficult task since both are very similar. Another one is that individual clothing items display many different appearance characteristics. For example, shirts have a wide range of appearances based on cut, color, material and pattern. Occlusions from other humans or objects, viewing angle and heavy clutter in the background further complicates the problem.

The work developed in this M.Sc. dissertation [7] has contributed to address the aforementioned challenges. The following contributions may be observed of this dissertation:

- **A new dataset for image annotation have been introduced [8].** The dataset is composed of images, tags and comments crawled from two fashion-related social networks, namely pose.com and chictopia.com. The whole dataset is composed of approximately five thousands images and a set of 31 classes.
- **A set of experiments was conducted to evaluate different visual feature representation and to analyze the best configuration for each type in the context of clothing annotation [8].** Ten global (color, shape and texture) and six local descriptors (associated with two mid-level techniques) were evaluated for the clothing annotation task.
- **Two different methods for clothing annotation that exploits association rules to create the classifiers [9].** The first approach, called Multi-modal and Multi-label Clothing Annotation algorithm (MMCA), which uses a feature vector of each single image associated with its labels as an instance. The second one, Multi-label, Multi-

TABLE I
DATASETS.

	pose.com	chictopia.com
Number of photos	2,306	1,579
Number of tags	7,501	5,093
Tags per photo	3.25	3.23

modal and Multi-instance Clothing Annotation method (M3CA), which uses pair of images as instances.

- **Novel multi-scale clothing parsing algorithm using Convolutional Networks.** Specifically, we propose a Multi-scale Convolutional Network (M-CNN), that creates a hierarchy of networks, where the first level processes a large amount of images with bigger dimension while the last one handles just a small amount of tiles with tiny size. This multi-scale strategy allows the method to capture minimal details of each image contributing to a more robust parsing algorithm.

II. Proposed Dataset

The proposed dataset is composed of images and associated tags crawled from two fashion-related social networks, namely pose.com and chictopia.com. Basic information about the resulting datasets is shown in Table I. The whole dataset for our realistic scenario is composed of approximately five thousands images. Combining labels from both datasets leads us to a set of 31 discrete possibilities, including "bag", "bathing suit", "belt", "booties", "cape", "coat", "dress", "glass", "gloves", "hat", "headband", "jacket", "jewelry", "jumpsuit", "pants", "pumps", "sandals", "scarf", "shirt", "shoes", "shorts", "skirt", "sneakers", "socks", "suit", "sweater", "tights", "umbrella", "underwear", "vest" and "wallet".

III. Descriptor Evaluation

A myriad of visual descriptors technique have been proposed and used in the literature achieving satisfactory results in various applications, but many of them have never been used in clothing identification tasks. Furthermore, different descriptors may produce different results depending on the data, it is imperative to design and evaluate many descriptor algorithms in order to find the most suitable ones for each application [10] Also, distinct descriptors may provide complementary information about images, so the combination of multiple descriptors is likely to provide improved performance when compared with a descriptor in isolation. However, the optimal combination of descriptors is data-dependent, as well as a hard task depending on the problem, since different descriptors may produce different results.

The contribution published in [8] presents an evaluation of ten global descriptors [11], [12], [13], [14], [15], [16], [17], [18], [19] that encode color, shape and texture properties for clothing annotation task. To evaluate these descriptors, we proposed a methodology based on the Lazy Association Classifier (LAC) [20]. Amongst interesting conclusions, we could pointed out the SID descriptor [18] is the best one amongst all of them.

Another contribution in this context, which was published in [9], is the analysis of two mid-level representations [21], [22] considering six local descriptors [23], [24], [25], [26], [27], [28]. This evaluation was performed considering several possible parameters, such as dictionary size. However, mid-level approaches were not so effective for this kind of application given the problem to separate the foreground from the background of the image.

IV. Clothing Annotation Techniques

Clothing annotation is a task that may be described as assigning short textual descriptors or keywords (called tags) to images. These tags are related to specific garment items, such as shirts, trousers and shoes, and multiple tags may be associated with an arbitrary image. We formulate this task as a supervised classification problem: a process that automatically builds a classifier from a set of previously labeled/annotated examples (i.e., the training-set). Then, given an arbitrary image (i.e., an image in the test-set), the classifier recognizes the labels/tags that are more likely to be associated with it. Figure 1 presents a overview of the proposed methods.

The contributions published in [8], [9] presents two novel approach for clothing parsing of everyday photos. Specifically, first, we propose a Multi-modal and Multi-label Clothing Annotation algorithm (MMCA), presented in Figure 3, that uses the pointwise approach, which is the most commonly used strategy [29]. According to [30], the pointwise approach employs the feature vector of each single image as an instance. In this case, each instance in the training set is composed of the visual and textual features (labels) of an image, while the test set is only composed by the visual features of an image. Second, we propose a Multi-label, Multi-modal and Multi-instance Clothing Annotation method (M3CA), presented in Figure 2, which is based on the pairwise approach, which is usually defined as an input space that represents instances as being a pair of images, both represented as feature vectors [30]. Hence, each data instance, in the training and in the test set, is a pair of images: the query image and the base image. Labels associated with base image are always known in advance in all sets (i.e., base labels) and labels associated with the query image are only known in advance in the training set (i.e., query labels). So, the only difference between the training and the test set is the query labels that are only known in the the former. Thus, for the training set, each instance is composed of a set of base and query labels, plus a set of distances between the images and, while for the test set, each instance is composed by only the base labels and the visual distances between the images.

Both methods use classifiers composed of association rules [31], which are essentially local mappings $X \rightarrow y$ relating a combination of features in instance X to a label y. These rules are used collectively, resulting in a membership probability for each label. In order to provide fast learning

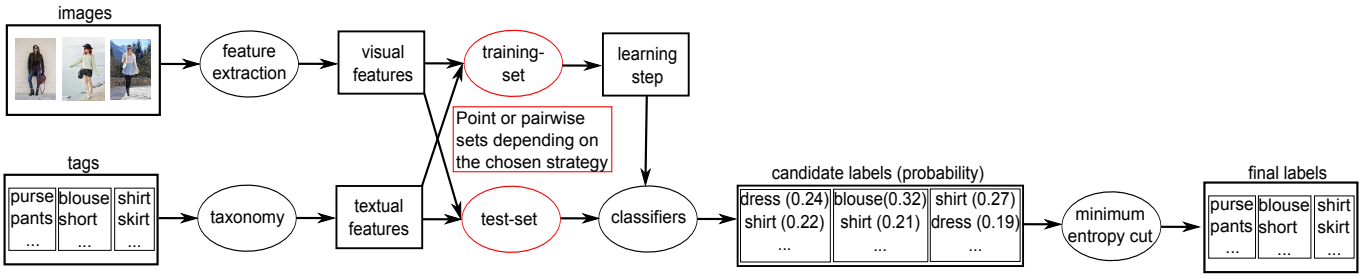

Fig. 1. Whole pipeline of the proposed methods for clothing annotation.

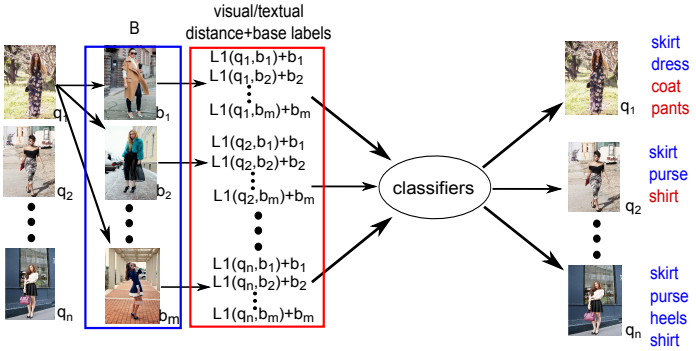

Fig. 2. Illustration of pairwise approach. In this case, the classifiers are already trained with paired images as well. Predicted labels in blue represent right labels while red ones represent wrong predictions.

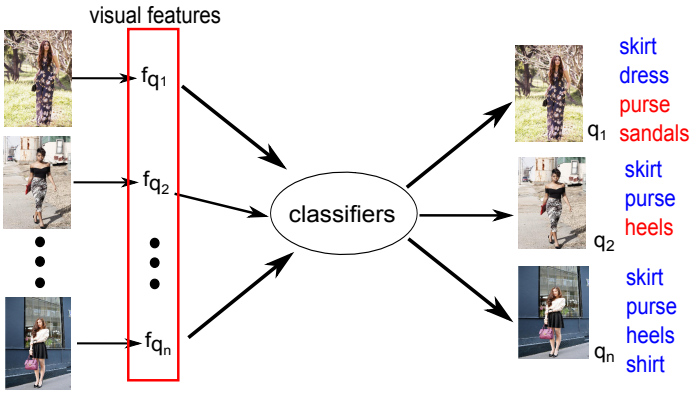

Fig. 3. Illustration of pointwise approach. Predicted labels in blue represent right labels while red ones represent wrong predictions.

times, the proposed algorithm extracts rules on a demand-driven basis – instead of learning a single and potentially large classifier which could be applicable to all instances in the test-set, our algorithm builds multiple small classifiers, one for each instance in the test-set. Typical solutions to multi-label classification employ the top-k approach [32], where a pre-determined threshold k is used to select the labels to be assigned to the query image. That is, only the k labels with the highest membership probabilities are assigned. Instead of relying on this parameter, we propose an entropy-minimization multi-instance approach which finds a different cut point for each instance in the test-set.

Furthermore, for the MMCA algorithm, we also proposed

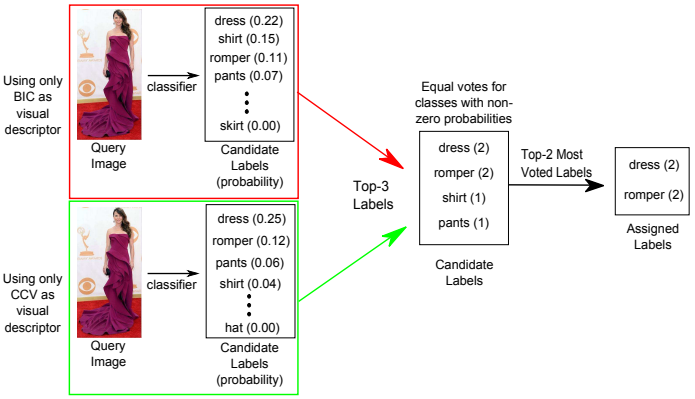

Fig. 4. Illustration of a proposed combination of the MMCA approach considering only BIC and CCV descriptors: the majority voting consider each class, with probability more than zero, as a vote with equal weight. A top-k defines which labels should be assigned.

two combination methods in order to join classifiers that use different visual features looking for improvements in the overall accuracy. The proposed algorithms, published in [9], may appear very similar to some ensemble methods in the literature, like bootstrap aggregating or bagging, but they differ from them because: (i) the classifiers are trained with different features (ii) the training set used is always the same for every classifiers (only the features used are different), and (iii) the misclassification of a classifier is never used again. First combination method, called Majority Voting (MV), gives each candidate label the same weight when voting. More specifically, for each instance a classifier generates, as presented, a ranking with the labels and its probability. This ranking is pruned using a top-k approach, and then, each remaining label (the ones with higher probability) gives an equal vote, creating a final ranking ordered by the votes. This final ranking is pruned again (also using a top-k method), resulting in the final set of labels that is assigned to the image. Figure 4 presents a example of this method considering classifiers trained using BIC and CCV visual descriptors. The second proposed combination method, called Majority Probability (MP), gives each candidate label a weight (equal its probability) when voting. Specifically, for an instance, the method generates a final ranking by calculating the mean probability of each label considering all the rankings. Then, the final rank is pruned in top-k way. Figure 5 presents a

example of this method considering classifiers trained using BIC and CCV visual descriptors.

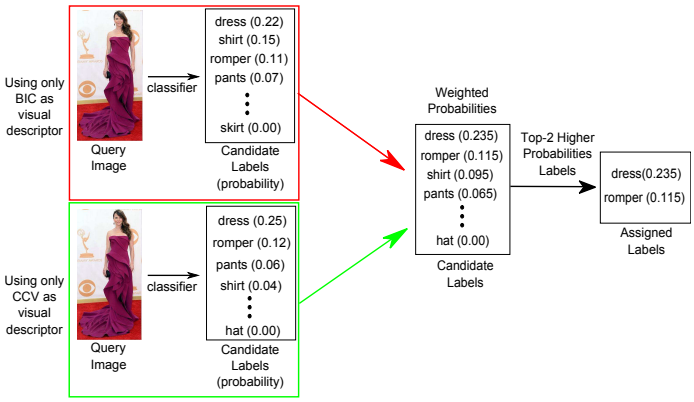

Fig. 5. Illustration of a proposed combination of the MMCA approach considering only BIC and CCV descriptors: the majority probability calculates the mean of all labels and a top-k is used to define which labels should be assigned.

The comparison between the proposed methods and some baselines are presented in Figure 6. The α parameter is the size of mask used to determine the region of interest, i.e., the place where the person might be. Please refer to [7] for a more detailed discussion. For both dataset, M3CA provides accuracy improvements that vary from 20% (M3LDA top-3) to 30% (M3LDA top-7). The combination of MMCA yields better accuracy than the M3CA approach, however, the MMCA approach, without combination, was not capable to achieve accuracy close to the M3CA method. Despite of achieving best results, the accuracy of the M3CA are almost as good as the combinations, but with much less processing time, since to get the combination, we need to get all results from each descriptor.

V. Clothing Parsing Techniques

Image parsing may be described as a process of partitioning an image into multiple segments (sets of pixels) in order to simplify its representation into something that is more meaningful and easier to analyze. In this case, these sets correspond to specific garment items in the image. We formulate this task using a Multi-scale Convolutional Network, or simply M-CNN, presented in Figure 7, that creates a hierarchy of networks, where the first level processes a large amount of images with bigger dimension while the last one handles just a small amount of tiles with tiny size. This multi-scale strategy allows the method to capture minimal details of each image contributing to a more robust parsing algorithm. To define which images go from one level to another, a entropy strategy was applied.

The entropy [33], a measure commonly used in information theory, characterizes the (im)purity of an arbitrary collection of examples. In this case, it denotes the purity of a single patch in relation to the number of classes associated to it, i.e., the more classes related to the patch the higher entropy it has (more impure). As introduced, entropy helps our approach to defined which patches are considered classified and which ones are not.

TABLE II
Clothing Parsing results.

Method	Pixel Accuracy (%)
Pointwise+BoW+SIFT+Rule Size 2	24.45
M-CNN	**40.79**

Specifically, the proposed method uses, in this case, three different network levels[1] which process images with different granularities, i.e., after every level the images are decomposed into smaller patches, allowing the network to capture minimal details. In the first level, larger images are processed in a robust network. Images with low entropy already get their final class in this level, while the others with high entropy (classification still undefined) are split into smaller patches and go to the next one. Remaining images without classification are again divided into even smaller patches and, finally, classified in the third level. At the end, we have a class associated with each patch of the image and a segmentation mask may be built.

Some obtained results are presented in Table II based on the overall accuracy. The pointwise approach was used as baseline. It is possible to see that the pointwise approach for clothing parsing achieve better result than this same method for clothing annotation, since for the former, there is no effect of the background. However, the M-CNN approach achieved much better results than the pointwise one, verifying that the proposed method is very promising.

VI. Conclusions and Future Work

In this Msc dissertation, methods to solve clothing annotation and parsing were proposed. We address several challenges, such as differ similar types of garment items and create pattern for a cloth since it may appears with different cut, color, material and pattern. It was completed in two years (from February 2013 to March 2015) and has resulted in one international journal papers [9] as well as the same number of conference paper [8].

Considering the clothing annotation, future work includes trying different classifiers, using deep features (ones extracted from pre-trained neural networks) and using Markov and Conditional Random Fields to include context. For the clothing parsing, future works includes improving the Convolutional network to use more adaptable inputs (instead of fixed-size ones), adapting the network to allow variable number of layers (in this work, we only use three) and using a different method at the final layers of the method, such as a fuzzy one.

Acknowledgment

The authors would like to acknowledge grants from CNPq, CAPES, Fundação de Apoio à Pesquisa do Estado de Minas Gerais (Fapemig), PRPq/Universidade Federal de Minas

[1]For this application, only three network levels were used because of the relative small size of the image and the benefit between patch size and processing time.

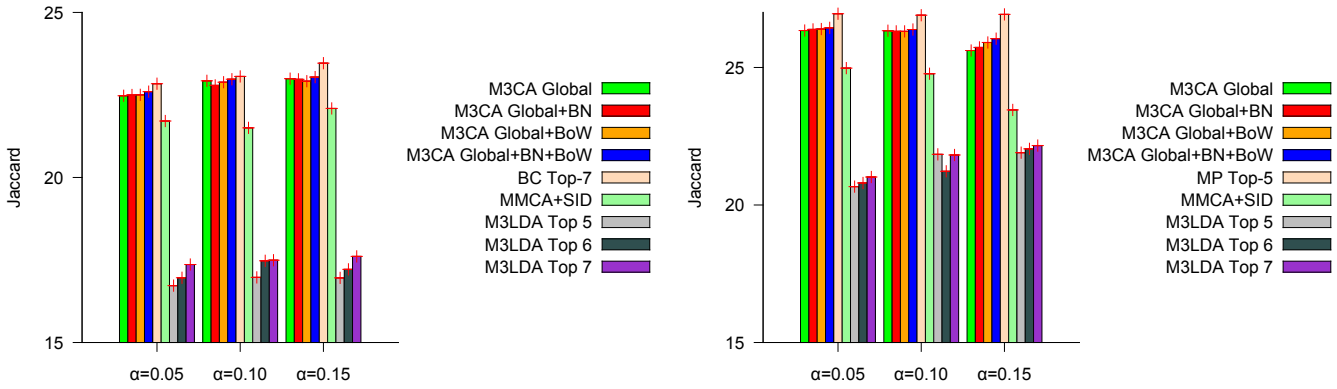

Fig. 6. The results of the M3CA and the baseline for Chictopia (left) and Pose (right). We also considered the best MMCA using SID, and the best combination algorithm for each dataset.

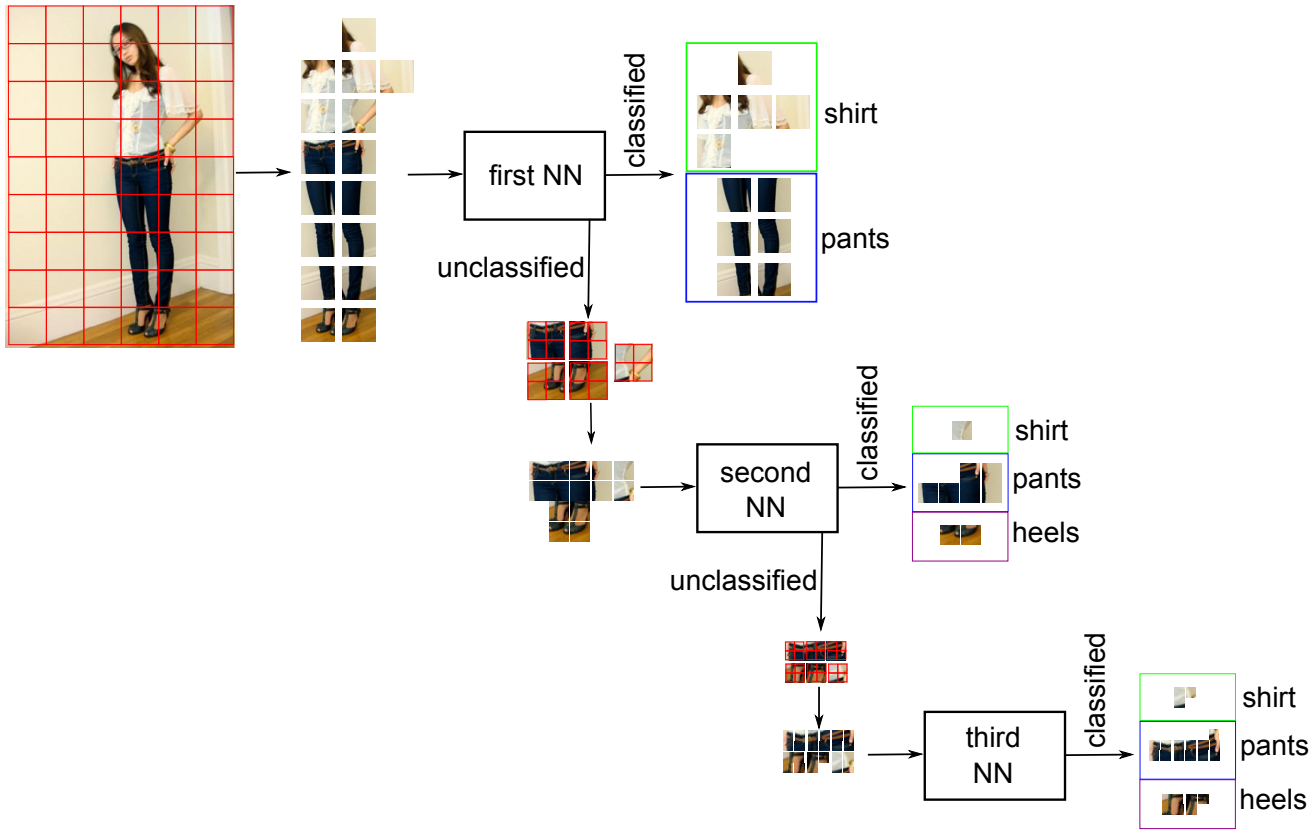

Fig. 7. Overview of the M-CNN approach. The original image is split into little tiles that are candidates to be classified in the first level network. The unclassified tiles are split again and goes for the second network. The same occurs on the last level of our architecture.

Gerais, Finep, and InWeb – the Brazilian National Institute of Science and Technology for the Web.

References

[1] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3570–3577.

[2] M. Weber, M. Bäuml, and R. Stiefelhagen, "Part-based clothing segmentation for person retrieval," in *8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2011*, 2011, pp. 361–366.

[3] A. C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Conference on Computer Vision and Pattern Recognition*, 2008.

[4] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos." in *International Conference on Image Processing*, 2011, pp. 2937–2940.

[5] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3330–3337.

[6] Y. Kalantidis, L. Kennedy, and L. Li, "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos," in *International Conference on Multimedia Retrieval*,

2013, pp. 105–112.

[7] K. Nogueira, "Statistical and deep learning algorithms for annotating and parsing clothing items in fashion photographs," dissertation, Universidade Federal de Minas Gerais, 2015.

[8] K. Nogueira, A. A. Veloso, and J. A. dos Santos, "Learning to annotate clothes in everyday photos: Multi-modal, multi-label, multi-instance approach," in *27th Conference on Graphics, Patterns and Images, SIBGRAPI 2014*. IEEE Computer Society, 2014, pp. 327–334.

[9] ——, "Pointwise and pairwise clothing annotation: combining features from social media," *Multimedia Tools and Applications*, vol. 75, no. 7, pp. 4083–4113, 2016.

[10] J. dos Santos, O. Penatti, P. Gosselin, A. Falcao, S. Philipp-Foliguet, and R. Torres, "Efficient and effective hierarchical feature propagation," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. PP, no. 99, pp. 1–12, 2014.

[11] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *IEEE 10th Conference on Computer Vision and Pattern Recognition, (CVPR 1997*, 1997, pp. 762–768.

[12] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *Proceedings of the 2002 ACM International Conference on Information and Knowledge Management, CIKM 2002*, 2002, pp. 102–109.

[13] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proceedings of the 4th ACM International Conference on Multimedia, ICM 1996*, 1996, pp. 65–73.

[14] F. Mahmoudi, J. Shanbehzadeh, A. Eftekhari-Moghadam, and H. Soltanian-Zadeh, "Image retrieval based on shape similarity by edge orientation autocorrelogram," *Pattern Recognition*, vol. 36, no. 8, pp. 1725–1736, 2003.

[15] B. Tao and B. W. Dickinson, "Texture recognition and image retrieval using gradient indexing," *Journal of Visual Communication and Image Representation*, vol. 11, no. 3, pp. 327–342, 2000.

[16] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[17] C. Huang and Q. Liu, "An orientation independent texture descriptor for image retireval," in *The 5th IEEE International Conference on Computer and Computational Sciences, ICCCS 2007*, 2007, pp. 772–776.

[18] J. Zegarra, N. Leite, and R. Torres, "Wavelet-based feature extraction for fingerprint image retrieval," *Journal of Computational and Applied Mathematics*, 2008.

[19] M. Unser, "Sum and difference histograms for texture classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 1, pp. 118–125, 1986.

[20] A. Veloso, W. Meira Jr, and M. J. Zaki, "Lazy associative classification," in *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 2006, pp. 645–654.

[21] S. E. F. de Avila, N. Thome, M. Cord, E. Valle, and A. de Albuquerque Araújo, "BOSSA: extended bow formalism for image classification," in *18th IEEE International Conference on Image Processing, ICIP 2011*, 2011, pp. 2909–2912.

[22] J. Sivic and A. Zisserman, "Video google: Efficient visual search of videos," in *Toward Category-Level Object Recognition*, 2006, pp. 127–144.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[24] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[25] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: binary robust independent elementary features," in *European Conference on Computer Vision*, 2010, pp. 778–792.

[26] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: binary robust invariant scalable keypoints," in *International Conference on Computer Vision*, 2011, pp. 2548–2555.

[27] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *International Conference on Computer Vision*, 2011, pp. 2564–2571.

[28] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: fast retina keypoint," in *Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2012, pp. 510–517.

[29] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, no. 1, pp. 346–362, 2012.

[30] T. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[31] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM International Conference on Management of Data, SIGMOD 1993*, 1993, pp. 207–216.

[32] A. Veloso, W. M. Jr., M. A. Gonçalves, and M. J. Zaki, "Multi-label lazy associative classification," in *11th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2007*, 2007, pp. 605–612.

[33] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. MIT Press, 2010.