Pointwise and pairwise clothing annotation: combining features from social media

Article *in* Multimedia Tools and Applications · April 2016

DOI: 10.1007/s11042-015-3087-2

CITATIONS
10

READS
330

3 authors:

Keiller Nogueira
University of Liverpool
57 PUBLICATIONS **2,662** CITATIONS
SEE PROFILE

Adriano Veloso
Federal University of Minas Gerais
220 PUBLICATIONS **4,528** CITATIONS
SEE PROFILE

Jefersson A. dos Santos
Federal University of Minas Gerais
141 PUBLICATIONS **4,192** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

Pointwise and Pairwise Clothing Annotation: Combining Features from Social Media

Keiller Nogueira* · **Adriano Alonso Veloso** · **Jefersson Alex dos Santos**

the date of receipt and acceptance should be inserted later

Abstract In this paper, we present effective algorithms to automatically annotate clothes from social media data, such as Facebook and Instagram. Clothing annotation can be informally stated as recognizing, as accurately as possible, the garment items appearing in the query photo. This task brings huge opportunities for recommender and e-commerce systems, such as capturing new fashion trends based on which clothes have been used more recently. It also poses interesting challenges for existing vision and recognition algorithms, such as distinguishing between similar but different types of clothes or identifying a pattern of a cloth even if it has different colors and shapes. We formulate the annotation task as a multi-label and multi-modal classification problem: (i) both image and textual content (i.e., tags about the image) are available for learning classifiers, (ii) the classifiers must recognize a set of labels (i.e., a set of garment items), and (iii) the decision on which labels to assign to the query photo comes from a set of instances that is used to build a function, which separates labels that should be assigned to the query photo, from those that should not be assigned. Using this configuration, we propose two approaches: (i) the pointwise one, called MMCA, which receives a single image as input, and (ii) a multi-instance classification, called M3CA, also known as pairwise approach, which uses pair of images to create the classifiers. We conducted a systematic evaluation of the proposed algorithms using everyday photos collected from two major fashion-related social media, namely `pose.com` and `chictopia.com`. Our results show that the proposed approaches provide improvements when compared to popular first choice multi-label, multi-modal, multi-instance algorithms that range from 20% to 30% in terms of accuracy.

Department of Computer Science, Universidade Federal de Minas Gerais (UFMG)
Avenida Presidente Antônio Carlos, 6627,
Belo Horizonte, MG, Brazil – CEP 31270-901
{keiller.nogueira,adrianov,jefersson}@dcc.ufmg.br
*Corresponding author - Phone/fax: +55-31-3409-5860

Keywords Image annotation · Clothing Annotation · Bag of Visual Words · Machine Learning · Multi-Modal · Multi-Instance · Multi-Label

1 Introduction

Full image understanding, the main goal of visual recognition field, is still an open task composed by the tripod: annotation, segmentation and classification of an image. The first two tasks complement each other, since the former one recognizes objects of a scene, while the other recognizes and also locates them. Further, both tasks can be helpful to image classification. In this paper, we focus on the image annotation task, given its versatility and applicability in several applications or task, such as classification. Moreover, we focus on a specific application, in this case, the clothing annotation one.

Image annotation plays important roles in human pose estimation [70], action recognition, person search [68, 18], surveillance [72], cloth retrieval [28] and have applications in fashion industry [70]. Considering the last one, applications with fashion images gained a lot of visibility with the increase of social networks and the faster propagation 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/clothes and receive feedbacks about them. There are even specific networks for this, such as `pose.com` and `chictopia.com`. These social media channels carry a lot of information, such as tags and comments, that is propitious (and fundamental) to a better understanding of an image. When analyzed, this information may help retailers and e-commerce systems to capture new trends, helping to define new products and sales. To achieve this, it is 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 realized 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. However, only a very small percentage of images collected from social media has been associated with its clothing content [25], and manual methods, using expert or non-expert [51], are too expensive and maybe impracticable given the total amount of images. So, automatic annotation techniques appear as a very appealing alternative to reduce costs, but with difficult challenges to overcome, such as: (i) differ similar types of garment items (discern between a shirt and a coat, for example), (ii) different appearance characteristics (cut, color, material and pattern) for a same cloth, (iii) occlusions, (iv) viewing angle and, (v) cluttered background.

In this paper, we are particularly interested in the clothing annotation 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 labelled/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.

First, we propose a Multi-modal and Multi-label Clothing Annotation algorithm, or simply MMCA, that uses the pointwise approach [29], i.e., a feature vector of each single image as an instance. Second, we propose a Multi-label, Multi-modal and Multi-instance Clothing Annotation method, or just M3CA, based on the pairwise approach, which is usually defined as an input space that represents instances as being a pair of images (called, in this paper, query and base images), both represented as feature vectors [29] (in this case, composed of distances between the images). Hence, each data instance, in the training and test set, is a pair of images and the only difference between them is that the query labels (labels of the query image) are only known in the former set. Is it important to emphasize that both methods intend to exploit the similarity between images, since similar ones are likely to share common labels, and thus small distances are expected to increase the membership probabilities associated with the correct labels. As introduced, looking for improvements of the annotation results, the instances of both methods are composed of visual and textual features. To allow the methods to get all feasible information from the data, visual features were coded using different image content descriptors (global [24, 53, 39, 31, 57, 55, 23, 75, 63] and local [30, 6, 11, 26, 43, 2]), while textual ones were described with TF-IDF vectors.

Our classifiers are composed of association rules [1], 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 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. Instead of relying on top-k approaches [65] to select the labels that should be assigned to the query image, we propose an entropy-minimization multi-instance approach which finds a different cut point for each instance in the test-set.

The main contribution of this paper is the proposed framework for automatic clothing annotation. To the best of authors' knowledge, there is no framework capable to annotate clothes in so many different scenarios, such as multi-instance, multi-label and multi-modal. Also, it is not known a framework to annotate clothes using pairs of images, which allows the method to capture semantic information (difference between the images) related to each pair. In practice, we may observe the following detailed contributions of this work:

- Novel multi-instance, multi-label, multi-modal clothing annotation algorithms with the aggregation of different types of descriptors.
- Two different methods for clothing annotation that exploit association rules to create the classifiers: the MMCA (which follows a pointwise strategy) and the M3CA (which follows a pairwise strategy) approaches.
- A comparison between all proposed approaches which leads us to define the best one for our annotation task.
- An extensive 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.
- A systematic set of experiments, using a collection of everyday photos crawled from popular fashion-related social networks, reveals that our algorithm improves upon first choice learning algorithms [35], by a factor that ranges from 20% to 30% in terms of standard accuracy measures.

We organized the remainder of this paper in seven sections. Section 2 presents related work. Section 3 presents the background concepts necessary for the understanding of this work. Section 4 shows the details of the proposed approaches for the clothing annotation task. The evaluation protocol is detailed in Section 5. Experimental evaluation, as well as the effectiveness of the proposed algorithms, is discussed in Section 6. Finally, Section 7 concludes and points out future research directions.

2 Related Work

This section presents a review of the literature surrounding clothing annotation, as well as some works of image annotation, since the former is a subproblem of the latter. It is worth to mention that the focus of this work is the clothing annotation task, i.e, clothing parsing [48] and other clothing applications, such as fashionability [49], were not covered.

There has been a great effort in the last few years on the clothing recognition task, with some works focusing on the clothing annotation. This recent interest is, perhaps, boosted by fresh advances in pose estimation [74], which caused a lot of works to emerge [78]. Amongst these, the most popular works [77] combine supervised machine learning algorithms and visual feature extraction methods. However, there are approaches [37] that learn the annotations using the users feedback, also called implicit crowdsourcing. Considering this technique, some approaches towards automatic image annotation exploit multi-label models [27, 66], others employ multi-modal strategy to improve results [69] and few works have modelled the problem as a multi-instance problem [35]. Furthermore, there are works that combine these strategies looking for a better performance [35, 36].

As introduced, typically, in these applications, an image has more than one label associated with it. Thus, classifiers for this task considered as multi-label ones. Accordingly to [60], multi-label classification algorithms can be

categorized into two different groups: (i) problem transformation methods and, (ii) algorithm adaptation methods.

The first group includes methods that are algorithm independent, i.e., they transform the multi-label problem into one or more single-label problems. Usually, methods from this group tend to exploit probabilistic models, such as Bayesian or Gaussian ones, to generate adapted algorithms capable to handle and annotate different images. There is a lot of works in this group, which includes binary relevance method [32], binary pairwise classification [34] and label combination methods [42].

The second group includes methods that extend specific learning algorithms in order to handle multi-label data directly. Well-known approaches include Adaboost [27], decision trees [66], lazy methods [65, 71] and, more recently, neural networks [52]. The proposed approaches may be categorized in this group, since an adaptation of the learning algorithm was realized to allow the prediction of multiple labels.

In addition to multi-label classification, the multi-modal fusion, which is a new scenario created by combining multiple media data and their associated features, has gained a lot of attention recently [3] due to the benefit this aforementioned combination provides. Usually, this fusion of multiple modalities can provide complementary information and increase the overall accuracy of the task. There is a lot of feasible fusions, such as audio/video or video/textual, though the most common fusion, when working with images, is the visual/textual one. This fusion takes leverage of: (i) visual features, that come from the images (usually obtained with descriptors) and, (ii) textual ones, which may be simplified by the tags/comments associated with each image. Given the increasing amount of images that are currently available on the web with poor accuracy annotation, this fusion has become really popular since it is interesting to use this data to learn more accurate recognition models.

Accordingly to [3], multi-modal fusion algorithms can be categorized into three different groups: (i) feature level or early fusion [69], which combines the features extracted from the input data and then send as input to the classifiers, (ii) decision level or late fusion [21, 22], which isolates the features to create different combinations of classifiers using some criterion, and (iii) hybrid approach [35], which is a combination of both feature and decision level strategies, taking advantages of both.

Considering the early fusion method, Xie et al. [69] uses images weakly tagged to improve the image classification performance using statistical approaches. Using the late fusion strategy, Guillaumin et al. [21] proposed a work that combines visual and textual features, where the textual ones are represented by labels/tags associated with images crawled in social networks. Also, Guillaumin et al. [22] use image tags to improve the performance of the classifiers, but they do not assume their availability for test images. Considering the hybrid approach, Nguyen et al. [35] proposed a method where the fusion of multi-modalities may be performed in both decision level (labels) and feature level (visual/textual) by using different models. The proposed approaches are

classified into the late fusion strategy, since they combine visual and textual features from each images and deliver them to the learning algorithm, that uses this combination to create classifiers.

Aside the aforementioned types of classifiers, the multiple-instance learning is a variation of supervised learning, which is the task of learning classifiers from bags of instances [33] that may contain as many instances as possible. Recently, this kind of approach has become very popular for some specific problems because of the good results achieved. Between these works, [56] proposed a method that exploits a unified learning framework which combines the multiple-instance and single-instance representations for image annotation. Specifically, they use an integrated graph-based semi-supervised learning that associate these types of representations simultaneously. [17] proposed an improved Transductive Multi-Instance Multi-Label (TMIML) learning, which aims at taking full advantage of both labeled and unlabeled data to address the annotation problem. Both of these works also use the Corel5K dataset on their experiments.

More recently, [35] proposed a multi-label, multi-modal and multi-instance approach using Latent Dirichlet Allocation (M3LDA). Specifically, they build the gist of a scene using [38] algorithm and then, they consider each patch of image as an instance, what generated a myriad of items. Each instance may be represented by a bag of prototypes, which are obtained by clustering visual features of the patch [79]. Associating instances and tags, they built a learning algorithm based on Latent Dirichlet Allocation (LDA). With this approach, they can not only annotate the images as a whole but can also annotate its region, if possible.

Considering the fashion scenario, several works exploit one or more aforementioned techniques. [58] proposed a system, named "Virtual Stylist", which aims to help users to find out outfits that might fit them well. [54] proposed a multi-label clothing annotation approach that enables users to efficiently update the metadata interactively and incrementally. [46] introduced the recommendation of outfits for specific occasions based on textual input that defines the occasion and how the user wants to look like. More recently, the work of [67] described the recommendation of clothes based on the similarity between users and models appearing in fashion magazines while [25] presented a scalable approach to automatically suggest relevant clothing products, given a single image without metadata. They, actually, formulated the problem as cross-scenario retrieval where the query is a real-world image, while the products from online shopping catalogues are usually presented in a clean environment.

In this work, we propose clothing annotation techniques associating all aforementioned concepts. As mentioned, our methods combine textual features with visual ones, in a later fusion mode, looking for improvements. Furthermore, while most works [35] treat each region (keypoint) of an image as a instance to create a multi-instance classifier resulting in a myriad of features, our proposed pairwise method creates a classifier by pairing images and calculating the visual distance between them, which make the approach more

sparse and robust. A combination of local and global visual features allows the proposed method to leverage from both of them, exploiting the best of each one in our application. In addition, different from other works [56, 17], instead of using less realistic scenarios, our experiments were on full realistic ones using dataset crawled from the web with tags generated by users from around the globe.

3 Background Concepts

This section presents background techniques used in this work. The proposed clothing annotation methods exploit the traditional combination of machine learning methods and visual image features. Thus, to extract the visual elements of the images, we used feature extraction algorithms (descriptors), which are categorized in low-level and mid-level ones.

Descriptors from the low-level work at extracting visual properties from the image via pixel-level operations, being crucial for any image analysis procedure. Two groups [47] of descriptors may be distinguished in this level: global and local. Global descriptors usually are simple and cheap to obtain since they rely on computing a representation that encodes global aspects of images. Local descriptors, in the other way, are more powerful to present object properties and are very precise, but more expensive to compute and compare, since a larger feature vector is produced. A local feature is an image pattern that differs (in some property) from its immediate neighborhood [62], such as points, edges or small patches. Typically, two types of patch-based approaches can be distinguished to extract local features [61]: (i) interest points, which are corners and blobs, and (ii) dense sampling, which are patches of fixed size over the whole image.

To extract all feasible information from the descriptors, mid-level feature extraction were created, which aims at transforming low-level descriptors into a global and richer image representation of intermediate complexity [9]. According to [9], in order to get the mid-level representation, the standard processing follows three steps: (i) low-level local feature extraction, (ii) (hard [41] or soft [20]) coding, which performs a transformation of the descriptors into a representation better adapted to the task, and (iii) (max or mean) pooling, which summarizes the coded features. Classification algorithms are then trained on the mid-level vectors obtained. Bag of Visual Words (BoVW) [50] and BossaNova [4] (special coding/pooling stage) are good examples of mid-level representations and were exploited in this work.

Considering these visual features, a myriad of machine learning methods could be exploited to create the classifiers, such as Support Vector Machines (SVM) and association rules. Despite all options, a machine learning method that could support multi-label, multi-modal and multi-instances strategies was desirable. Thus, Lazy Association Classifiers (LAC) [64] was chosen given its natural adaptation to multi-modal approaches, accepting visual and textual features without too much effort, and also because it permits the scalability

of the instances without increasing the processing time, since the number of classes is more relevant to the algorithm than the number of instances. In addition to this, the method easily allows the use of multi-label strategy since its output consists of a ranking with the classes and respectively probabilities.

LAC uses association rules [1] to produce classifiers that, depending on the task, may predict labels of an image or relevance related to a document. These rules are patterns describing implications of the form $X \xrightarrow{\theta} y_i$, where X is known as the antecedent of the rule while y_i is the consequent. The antecedent may be any combination of features, depending on the task, while the consequent may be any label or class. The rule does not express a classical logical application where X necessarily entails y_i. Instead it denotes the tendency of observing y_i when X is observed. Each rule has a size, which defines how many terms, considering antecedent and consequent, it has. Although the consequent always has one term (since each rule predicts exactly one class/label), the antecedent may have more than one term. For instance, a rule with size two has only one term in the antecedent and consequent while one with size three may have two elements in the antecedent and one in the consequent. The strength of the association between the antecedent and the consequent is measured by a statistic θ, which is known as confidence [1] and is simply the conditional probability of the consequent given the antecedent.

From a labelled training-set $\mathcal{D}$, the algorithm extracts **all possible rules** creating a set $\mathcal{R}$ of rules composed of rules used to predict classes/labels $\mathcal{L}_X$ that approximates as accurately as possible $\mathcal{L}_X^*$, which represents the ground-truth of the instance. It is important to emphasize that different sets may be generated from different labelled training-sets, i.e., if a training-set has different labels, different rules are generated. However, in this paper, the training-set has basically the same labels, thus the same rules are generated and there is no better or worse training-set. Each $\{X \rightarrow y_i\} \in \mathcal{R}$ is a vote given for label/class y_i. Thus, given an instance Z in test-set $\mathcal{T}$, a rule is a valid vote if it is applicable to Z, i.e., a rule $\{X \rightarrow y_i\}$ is said to be applicable to instance $Z \in \mathcal{T}$ if all intervals in X are in Z. From these applicable rules, a subset $\mathcal{R}_Z^{y_i}$ may be create. It corresponds to rules applicable to a specific instance Z predicting specific class/label y_i. Thus, a score for y_i is given by averaging the votes (weighted by confidence θ) in $\mathcal{R}_Z^{y_i}$, as shown in Equation 1. The likelihood $\hat{p}(y_i|Z)$ of an instance Z being associated with class/label y_i is obtained by normalizing the scores (to restrict the sum to exactly one). At the end, for each instance, LAC generates a ranking with all the classes/labels associated with its likelihood (probability).

$$s(Z, y_i) = \frac{\sum \theta(X \rightarrow y_i)}{|\mathcal{R}_Z^{y_i}|} \tag{1}$$

where $X \subseteq Z$ and $|\mathcal{R}|$ represents the set size.

4 Machine Learning Approaches for Clothing Annotation

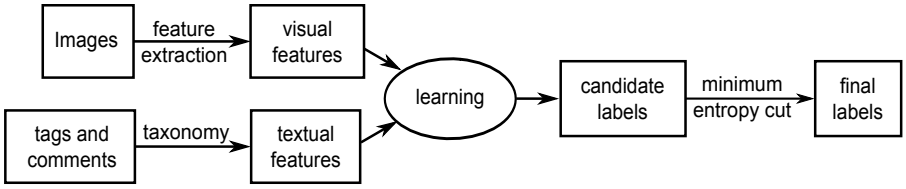


Fig. 1: An overview of the proposed methods.

In this section, we present the proposed algorithms for automatic clothing annotation. Figure 1 presents an overview of the proposed pipeline, which can be basically described in three steps: (i) visual and textual feature extraction, (ii) learning process, which is responsible to create the membership probability for each label, and (iii) labels selection, which uses an entropy-minimization algorithm to select labels to be assigned to the image.

Specifically, visual features were extracted using several types of descriptors, while textual ones were generated from tags and comments by using TF-IDF, which presents successful results in multi-modal approaches in the literature [14]. After extracting the features, there are two possible approaches based on Lazy Associative Classifiers (LAC) algorithm, presented in Section 3, which was chosen as learning algorithm given its natural adaptation to multi-modal, multi-label and multi-instance approaches: (i) pointwise, presented in Section 4.1 and, (ii) pairwise, presented in Section 4.2. Some methods proposed to combine the pointwise approach results are presented in Section 4.1.2. Figure 2 shows an overview of the former one, where an input consists of a single image, while Figure 3 presents the latter approach, where pairs of images are used as input. To create classifiers capable of associating similar clothes and then annotate images, both methods exploit usual similarities of social networks, such as images sharing common garment items are likely to share similar visual elements (e.g., color, texture and shape) or people tend to use similar tags/comments with images that share common garment items. As introduced, both algorithms build classifiers on a demand-driven basis using LAC and each classifier returns membership probabilities for each label. Finally, a set of predicted labels is generated by minimizing the entropy of the membership probabilities returned by LAC. The Minimum Description Length (MDL) algorithm is used to define which labels should be assigned to the query image and is presented in Section 4.1.1. Although being used in both proposed methods, to simplify its complexity, the formalism of the MDL algorithm is introduced considering only the pointwise method.

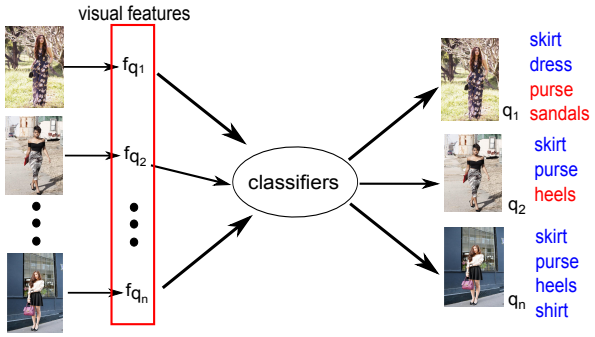


Fig. 2: The proposed pointwise approach. Predicted labels in blue represent right labels while red ones represent wrong predictions.

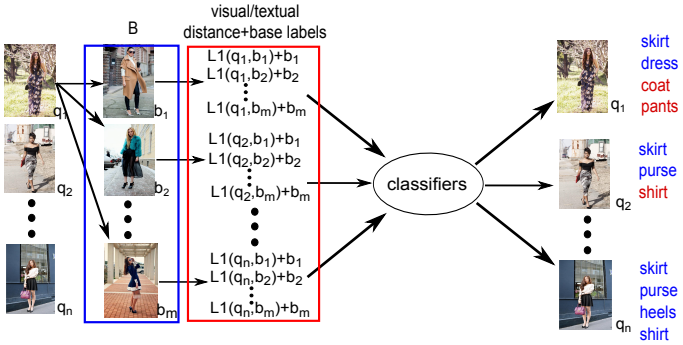


Fig. 3: The proposed pairwise approach. In this case, the classifiers are already trained with paired images as well. Predicted labels in blue represent right labels while red ones represent wrong predictions.

4.1 Pointwise Approach

In this approach, the pointwise method for automatic clothing annotation, named Multi-modal/Multi-Label Clothing Annotation (MMCA), is presented. In this case, a set of single image features (visual and textual) is provided as input [29]. Formally, Definition 1 describes the input of the MMCA approach.

Definition 1 A **pointwise instance** is composed by an image q associated with its visual and textual features. Specifically, a pointwise instance is represented by a feature vector $(\tilde{q}) = \{f_1, f_2, \cdots, f_m, v_1, v_2, \cdots, v_n\}$, where f represents the visual feature vector (of size m) and v represents the n textual features (labels).

As introduced, the proposed method uses association rules [1], which are called, in this case, garment rules, since each rule predict garment items. These rules are composed of an antecedent and a consequent, as presented in Section 3, which are represented, in this case, by any mixture of visual and textual

features and a label l_i (i.e., a garment item), respectively. Definition 2 formally presents a generic garment rule.

Definition 2 A **garment rule** has the following form:

$$\overbrace{\{\ f_j \wedge \ldots \wedge f_z \wedge v_t \wedge \ldots \wedge v_u\ \}}^{\text{Distance intervals}} \xrightarrow{\theta} l_i \begin{cases} \text{“trousers”}, \\ \text{“skirt”}, \\ \text{“handbag”}, \\ \text{etc.} \end{cases}$$

where $j \geq 1$ and $z \leq m$, and $t \geq 1$ and $u \leq n$. f represents the visual feature vector while v represents the textual ones. l_i is label assigned to this rule while θ is the confidence, as aforementioned. The operator “$\wedge$” represents that the antecedent of a rule is formed with the simple presence of a determined combination of features and labels. These combinations work like a signature to the rule.

The algorithm receives as input a labelled training-set $\mathcal{D}$ composed of pointwise instances, as in Definition 1. Distances between these instances are discretized [16] and then assigned to distance intervals[1], in order to allow for the enumeration of garment rules. The test-set $\mathcal{T}$ also consists of records of the same form, except that labels are unknown. As presented, from each pointwise instance $\tilde{q} \in \mathcal{D}$, the algorithm extracts a rule-set $\mathcal{R}$ composed of garment rules used to predict labels $\mathcal{L}_q$, which approximates, as accurately as possible, the ground-truth $\mathcal{L}_q^*$ of the instance $\tilde{q}$, i.e., the garment items associated with image q. As introduced, considering $\{\tilde{q} \rightarrow l_i\} \in \mathcal{R}$, it is possible to extract a subset $\mathcal{R}_{\tilde{z}}^{l_i}$ composed of rules applicable to $\tilde{z} \in \mathcal{T}$ (all intervals in $\tilde{q}$ are in $\tilde{z}$) predicting label l_i. This subset $\mathcal{R}_{\tilde{z}}^{l_i}$ may be averaged generating a score for label l_i, as shown in Equation 1. Finally, the likelihood $\hat{p}(l_i|\tilde{z})$ of an instance $\tilde{z}$ being associated with label l_i is obtained by normalizing the scores generating a ranking with the labels and its probability for each instance.

4.1.1 Minimum Description Length

The Minimum Description Length approach, or simply MDL, is used to induce two partitions, for each instance, over the space of membership probabilities returned from the classifier. Based on the entropy, a partition is selected and, then, assigned to the instance. This approach is more robust and adaptable when compared to the top-k method, typically used on multi-label problems [65], since it may vary the number of assigned labels based on the training instances. More specifically, given an instance $\tilde{q}$ and a set of candidate labels $\mathcal{L}_{\tilde{q}}$ provided by the classifier,[2] we must find a cut point $c_{\tilde{q}}$ which delimits labels that are likely to be associated with the query image from those that

[1] Hereafter we refer each f_i as the corresponding interval.

[2] Labels for which $\hat{p}(l_i|\tilde{q}) > 0$.

are not. In other words, we must find a threshold $c_{\tilde{q}}$, so that only labels in $\mathcal{L}_{\tilde{q}}$ for which $\hat{p}(l_i|\tilde{q}) > c_{\tilde{q}}$ are finally predicted.

As introduced, MDL searches for a threshold $c_{\tilde{q}}$ that provides the best entropy cut in the space induced by probabilities $\hat{p}(l_i|\tilde{q})\ \forall\ l_i \in \mathcal{L}_{\tilde{q}}$. Figure 4 illustrates the method. In the figure, symbol $\oplus$ indicates that the corresponding label l_i is associated with query image q. Similarly, symbol $\ominus$ indicates that the corresponding label l_i is not associated with query image q. Therefore, in the example, labels $\{l_4, l_5, l_6\}$ are associated with q (i.e., $\oplus$), while labels $\{l_1, l_2, l_3\}$ are not (i.e., $\ominus$). The figure shows three possible cut points for the instance, and the best entropy cut is exactly the one which minimizes the overall entropy in the probability space. Obviously, there are more difficult cases, for which it is not possible to obtain a perfect separation in the probability space, but the approach is general enough to handle such harder cases.

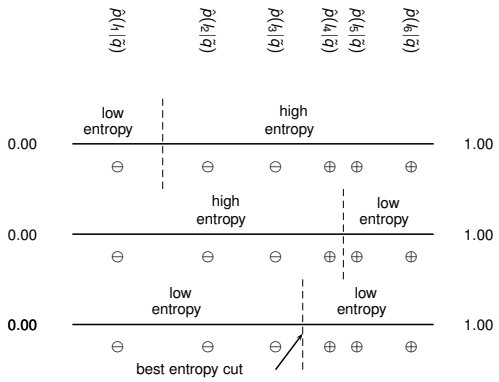


Fig. 4: Looking for the minimum entropy cut for a specific instance $\tilde{q}$.

Specifically, the idea is that any value of $c_{\tilde{q}}$ induces two partitions over the space of values for $\hat{p}(l_i|\tilde{q})$, i.e., one partition with probabilities that are lower than $c_{\tilde{q}}$, and another partition with probabilities higher than $c_{\tilde{q}}$. MDL seeks $c_{\tilde{q}}$ that minimizes the average entropy of these two partitions. Formally, consider a list $\mathcal{O} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_j \in \{\ominus, \oplus\}$ and y_j is a membership probability $\hat{p}(l_i|\tilde{q})$. This list is sorted such that $y_j \leq y_{j+1}$. Also consider c as a candidate value for $c_{\tilde{q}}$. In this case, $\mathcal{O}_c(\leq)$ is a sublist of $\mathcal{O}$ for which the condition $y_j \leq c$ holds for all $(x_j, y_j) \in \mathcal{O}_c(\leq)$. Similarly, $\mathcal{O}_c(>)$ is a sublist of $\mathcal{O}$ for which the condition $y_j > c$ holds for all $(x_j, y_j) \in \mathcal{O}_c(>)$. In other words, both $\mathcal{O}_c(\leq)$ and $\mathcal{O}_c(>)$ are partitions of $\mathcal{O}$ induced by c.

Firstly, the approach calculates the entropy in $\mathcal{O}$, as shown in Equation 2. Then, it calculates the sum of the entropies in each partition induced by c, according to Equation 3. Finally, it sets $c_{\tilde{q}}$ to the value of c that minimizes $E(\mathcal{O}) - E(\mathcal{O}_c)$.

$$E(\mathcal{O}) = -\left(\frac{\mathrm{N}_{\ominus}(\mathcal{O})}{|\mathcal{O}|} \times \log \frac{\mathrm{N}_{\ominus}(\mathcal{O})}{|\mathcal{O}|}\right) - \left(\frac{\mathrm{N}_{\oplus}(\mathcal{O})}{|\mathcal{O}|} \times \log \frac{\mathrm{N}_{\oplus}(\mathcal{O})}{|\mathcal{O}|}\right) \tag{2}$$

where $\mathrm{N}_{\ominus}$ gives the number of labels in $\mathcal{L}_{\tilde{q}}$ but not in $\mathcal{L}^*_{\tilde{q}}$, and $\mathrm{N}_{\oplus}$ gives the number of labels in $\mathcal{L}_{\tilde{q}}$ and also in $\mathcal{L}^*_{\tilde{q}}$.

$$E(\mathcal{O}_c) = \frac{|\mathcal{O}_c(\leq)|}{|\mathcal{O}|} \times E(\mathcal{O}_c(\leq)) + \frac{|\mathcal{O}_c(>)|}{|\mathcal{O}|} \times E(\mathcal{O}_c(>)) \tag{3}$$

To use the MDL approach, we employ a validation-set $\mathcal{V}$ composed of several instances $\tilde{q}$, so that both the true labels $\mathcal{L}^*_{\tilde{q}}$ and the predicted labels $\mathcal{L}_{\tilde{q}}$ are previously known for all instances in this set. Our goal is to build a function $\gamma(\mathcal{L}_{\tilde{q}})$ which receives as inputs a set of candidate labels $\mathcal{L}_{\tilde{q}}$ and returns the best entropy cut for these labels, predicting the labels. Thus, the function $\gamma(\mathcal{L}_{\tilde{q}})$ gives the mean of the best entropy cuts associated with instances $\tilde{q} \in \mathcal{V}$ having $\mathcal{L}_{\tilde{q}}$ as candidate labels. Equation 4 presents this mean. If there is no instances $\tilde{q} \in \mathcal{V}$ having specifically the candidate labels, then the function returns a mean of best cuts of all instances in the validation set.

$$\gamma(\mathcal{L}_{\tilde{q}}) = \frac{\sum c_{\tilde{q}}^{\mathcal{L}_{\tilde{q}}}}{\mathrm{N}_{\mathcal{L}_{\tilde{q}}}} \tag{4}$$

where $c_{\tilde{q}}^{\mathcal{L}_{\tilde{q}}}$ are best entropy cuts associated with the candidate labels $\mathcal{L}_{\tilde{q}}$ and $\mathrm{N}_{\mathcal{L}_{\tilde{q}}}$ is the number of validation instances associated with these labels.

4.1.2 Combination Methods Using MMCA

The combination methods proposed in this work join classifiers that use different visual features looking for improvements in the overall accuracy. The proposed algorithms may appear very similar to some ensemble methods in the literature, like bootstrap aggregating or bagging, but they are different since: (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 pointwise 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

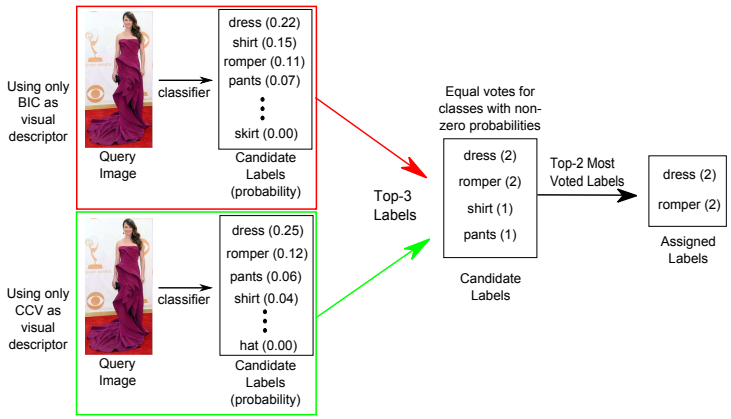


Fig. 5: Illustration of a proposed combination of the MMCA approach considering only BIC and CCV descriptors: the majority voting consider each class, with probability more than zero, as a vote with equal weight. A top-k defines which labels should be assigned.

a final ranking ordered by the votes. This final ranking is pruned again (also using a top-k strategy), resulting in the final set of labels that is assigned to the image. Figure 5 presents an 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 from all rankings. Then, the final rank is pruned in top-k way. Figure 6 presents an example of this method considering classifiers trained using BIC and CCV visual descriptors.

4.2 Pairwise Approach

In this approach, the pairwise method for automatic clothing annotation, named Multi-Modal/Multi-Label/Multi-Instance Clothing Annotation (M3CA), is presented. In traditional supervised learning, such as the proposed MMCA, an object is represented by an instance (usually, features) and associated with a class label. Although successful, some problems may not fit very well to this model, such as problems where the object may be associated with a multiple number of instances simultaneously, as for example, an image represented by a myriad of patches (feature vectors). To deal with this kind of problem, arise the multi-instance learning [29]. In this framework, an object is described by multiple instances. Formally, $\mathcal{X}$ represent the instance space and $\mathcal{Y}$ the set class labels. The task is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ from a given dataset $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, where $X_i \in \mathcal{X}$ is a set of instances $x_1^{(i)}, x_2^{(i)}, ..., x_{m_i}^{(i)}$, $x_j^{(i)} \in \mathcal{X}(j = 1, 2, ..., m_i)$, and $Y_i \in \mathcal{Y}$ is the set of

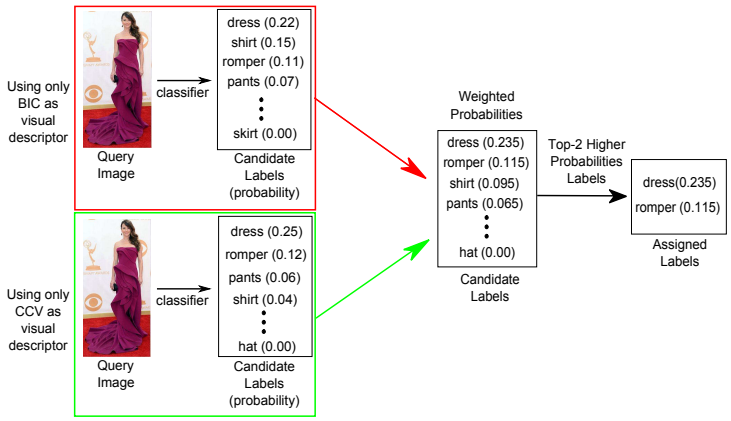


Fig. 6: Illustration of a proposed combination of the MMCA approach considering only BIC and CCV descriptors: the majority probability calculates the mean of all labels and a top-k is used to define which labels should be assigned.

labels $y_1^{(i)}, y_2^{(i)}, ..., y_{l_i}^{(i)}$, $y_k^{(i)} \in \mathcal{Y}(k = 1, 2, ..., l_i)$. In this case, we are also considering that a object may have more the one label. m_i represent the number of instances in X_i and l_i the number of labels in Y_i.

Therefore, this approach may be considered as a multi-instance one, since pairs of images are provided as input to our classification algorithm. In this case, a pair of images is denoted as an instance, as described in Definition 3.

Definition 3 A **pairwise instance** $(\widetilde{qb}) = (q, b)$ is composed by a base image b and a query image q. Labels associated with the base image b are called *base labels* and are always known in advance. Labels associated with the query image q must be predicted. This instance is represented by a set of (visual and textual) distances between q and b, along with the base labels. Specifically, a pairwise instance is represented as a list $(\widetilde{qb}) = (q, b) = \{f_1, f_2, \ldots, f_m, v_1, v_2, \ldots, v_n\}$, where f represents the distances between the visual feature vector (of size m) and v represents the n base labels.

It is important to highlight that the L1 distance function[3] was used to calculate the similarity between two images, since it is suitable for generate sparse vectors due to its property in producing results with zero or very small values.

The algorithm receives as input a labelled training-set $\mathcal{D}$ composed of records of the form $< q, B >$, where q is a query image and B is a bag of base images. The bag B is partitioned into multiple instances of the form $(\widetilde{qb}, \mathcal{L}_q^*) = ((q, b), \mathcal{L}_q^*)$, where $b \in B$, (q, b) is an instance as in Definition 3 and

[3] L1 distance function calculates the difference between two feature vectors by summing the absolute value of each keyword: $L1(P, Q) = \sum_{i=1}^{N} |p_i - q_i|$

$\mathcal{L}_q^*$ is a set of labels associated with the query image q (i.e., the garment items appearing in image q). Hence, in this case, it is known in advance the base and query labels, in addition to the feature distances between q and b. The test-set $\mathcal{T}$ also consists of records of the form $< q, B >$. Again, the bag B is partitioned into multiple instances $(\widetilde{qb}, ?) = ((q, b), ?)$. In this case, however, only the distances between images q and b and the base labels are known, whereas labels $\mathcal{L}_q^*$ are unknown and must be predicted. Hence, each data instance, in the training and test set, is a pair of images and the only difference between them is that the query labels are only known in the former set.

Just like the pointwise approach, the algorithm extracts a rule-set $\mathcal{R}$ composed of garment rules (as in Definition 2) from each pairwise instance $\widetilde{qb} \in \mathcal{D}$. As introduced, it is possible to extract, from $\mathcal{R}$, a subset $\mathcal{R}^{l_i}_{\widetilde{qb}}$ that contains rules applicable to specific pairwise instance $\widetilde{qb}$ predicting label l_i. As in Equation 1, a score for each label $s(\widetilde{qb}, l_i)$ is calculated using the confidence as weight. Finally, the likelihood $\hat{p}(l_i|\widetilde{qb})$ of an instance $\widetilde{qb}$ being associated with label l_i is obtained by normalizing the scores generating a ranking with the labels and its probability for each instance.

Analogously to the MMCA aproach, the M3CA algorithm also needs to build the function $\gamma(\mathcal{L}_{(\widetilde{qb})})$ to select the labels that should be assigned to the query image q. However, instead of using the function to directly predict the labels, as the MMCA approach, the M3CA needs to aggregate different pairwise instances related to a same query image q to, finally, predict the labels using the MDL algorithm. More specifically, a query image q may appear within several (i.e., n) pairwise instances $(\widetilde{qb}_i) = (q, b_i) \in \mathcal{T}$. For each instance $(\widetilde{qb}_i) = (q, b_i) \in \mathcal{T}$ a specific set of labels $\mathcal{L}_{(\widetilde{qb}_i)}$ is associated with q. The final set of predicted labels is given as $\mathcal{L}_q = \{\mathcal{L}_{(\widetilde{qb}_1)} \cup \mathcal{L}_{(\widetilde{qb}_2)} \cup \ldots \cup \mathcal{L}_{(\widetilde{qb}_n)}\}$. After aggregating the labels, we use the best entropy cut to predict the labels that should be associated with q.

5 Experimental Protocol

In this section, we present the experimental setup used in this work. Two scenarios were experimented:

1. Ideal scenario: consists of a small and manually segmented dataset, composed of 100 images, used to evaluate the visual descriptors and their best configuration (for example, the size of the visual dictionary). In this scenario, there is single-class classification (only one class should be assigned to each image) with 10 images per class. Therefore, the MMCA approach was used considering that the label with higher probability is assigned to the image.
2. Realistic scenario: consists of two datasets crawled from social networks used to analyze the proposed algorithms (MMCA and M3CA) and the baseline. Each image may have more than one label (garment items), which

Table 1: 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

makes this scenario a multi-label classification. The segmentation was realized automatically using a pose estimation algorithm, proposed by [74].

In this section, we distinguish some differences between the aforementioned scenarios. Section 5.1 presents some statistics of the datasets. The visual features are presented in Section 5.2 while Section 5.3 presents the textual ones. Section 5.4 presents the baselines used in this work. The experimental protocol used are presented in Section 5.5. Finally, Section 5.6 presents the measures used to evaluate the experiments.

5.1 Datasets

As introduced, the ideal scenario was designed to study the impact of the visual features over the overall accuracy in an attempt to avoid any external or unadvised error. This scenario consists of a dataset of 100 images (10 classes with 10 images per class) crawled from `instagram.com` between October 11 and November 10, 2013.

The realistic scenario was developed to evaluate the performance of the methods in a more real situation. Thus, we have crawled images and associated tags/comments from two fashion-related social networks, namely `pose.com` and `chictopia.com`. Basic information about the resulting datasets is presented in Table 1. The `pose.com` dataset was crawled from January 15, 2014 to January 25, 2014, which resulted in more than three thousand images. The `chictopia.com` dataset was crawled from January 25, 2014 to February 5, 2014 resulting in more than two thousands images. At the end, the whole dataset [4] for the realistic scenario is composed of, approximately, five thousands images. Combining labels from both datasets leads us to a set of 31 discrete possibilities: “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”.

Figure 8 shows the frequency of each label. As expected, some labels occur frequently (e.g., “coat”, “pants”, and “shirt”), while others occur only few times (e.g., “jumpsuit”, “suit” and “wallet”). Figure 7 shows the cumulative

[4] Both, Chictopia and Pose, datasets used in this paper are available for download at: `http://www.patreo.dcc.ufmg.br/downloads/fashion-datasets/`

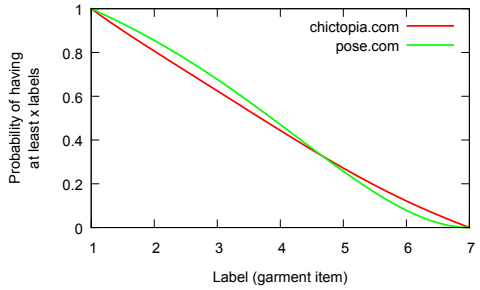
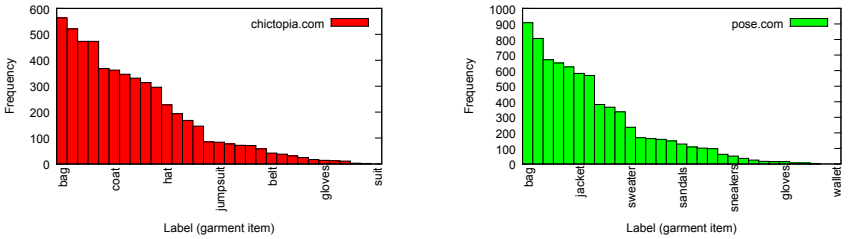


Fig. 7: Cumulative distribution function of labels for both datasets.

distribution function for labels in both datasets. The probability for an arbitrary image having at least x labels decreases almost linearly in both cases.



(a) (b)

Fig. 8: Frequency distribution related to the dataset: (a)-(b) for Chictopia and Pose, respectively.

When working with visual image descriptors, a pose estimation and image segmentation is needed since the background may make features more noise. Thus, in order to avoid the effect of background pixels over the description of the image, we have created a mask to separate relevant pixels, i.e., **a mask to delineate the human body**. For the ideal scenario, the mask creation was realized manually for each image. Figure 9 shows the original image and relevant pixels.

For the realistic scenario, we ran an human pose estimation algorithm [74] and then, create a mask based on the generated skeleton. Specifically, using the skeleton that estimates the human pose, we employ a factor of proportionality in order to enlarge each line of this estimation encompassing the entire pose and delineating the human body. So, we separate the pixels and obtain the final set of relevant pixels (i.e., non-background pixels). Figure 10 shows the original image, a pose estimation skeleton [74] and relevant pixels.

To exploit the benefits of this mask, an adaptation was necessary in visual descriptor algorithms to extract features only inside the region of interest

(a) (b)

Fig. 9: Example of manual segmentation: (a) Original image (b) Pixels of interest in white.

(mask). For the mid-level approaches, the mask was used as an intermediate step to select only relevant points before creating the visual dictionary, since a dense sampling was used.

(a) (b) (c)

Fig. 10: Example of automatic segmentation: (a) Original image (b) Skeleton generated by [74] (c) Pixels of interest in white.

We discarded images according to the proportion of background/non-background pixels. Specifically, we discarded all images for which the proportion of relevant pixels (non-background pixels) is lower than a fixed threshold α_{min}. We evaluate the impact of this parameter over the results in Section 6. Table 2 shows the number of remaining images (i.e., the final dataset) for different values of α_{min}.

5.2 Visual Descriptors

As presented in Section 3, there is a myriad of descriptors available in the literature [76] and choosing the most appropriate ones for a specific problem

Table 2: Images with enough relevant pixels.

α_{min}	# images `pose.com`	# images `chictopia.com`
0.05	1,308	1,257
0.10	969	937
0.15	578	421

is a hard task, since different descriptors may produce different results. In this paper, one contribution is to define the most interesting descriptors to solve the clothing annotation task. Ten global descriptors were selected and evaluated. This selection was based on extensive experiments performed in [44, 45, 40], which pointed out to some of the most interesting image descriptors in the current computer vision literature:

1. Color descriptors:
   (a) Auto-Correlogram Color (ACC) [24]
   (b) Border/Interior Pixel Classification (BIC) [53]
   (c) Color Coherence Vector (CCV) [39]
   (d) Global Color Histogram (GCH) [55]
   (e) Local Color Histogram (LCH) [55]
2. Texture descriptors:
   (a) Quantized Compound Change Hist. (QCCH) [23]
   (b) Local Activity Spectrum (LAS) [57]
   (c) Steerable Pyramid Decomposition (SID) [75]
   (d) Unser [63]
3. Shape descriptors:
   (a) Edge orientation auto-correlogram (EOAC) [31]

Six different feature extraction techniques were evaluated in order to select the best low-level descriptors to be used in the visual dictionary creation. These descriptors were chosen based on extensive experiments performed in [7]:

1. Scale-Invariant Feature Transform (SIFT) [30]
2. Speeded Up Robust Features (SURF) [6]
3. Binary Robust Independent Elementary Features (BRIEF) [11]
4. Binary Robust Invariant Scalable Keypoints (BRISK) [26]
5. Oriented FAST and Rotated BRIEF (ORB) [43]
6. Fast Retina KeyPoint (FREAK) [2]

For the BoVW approach, hard assignment followed by max polling were chosen at the coding and polling stage, respectively. This choices were made in order to get a more sparse histogram, that tends to be easier to learn with.

For BossaNova [4], localized soft assignment followed by BossaNova pooling were chosen at the coding and pooling stage, respectively. After creating the visual features, a two-step signature normalization (power-law and ℓ_2-normalization) is realized in order to get a more reliable histogram. It is important to highlight that, for all this techniques, we used a dense sampling to get the patches.

5.3 Textual Features

Textual features are represented by the tags and comments associated with an image, which may bring useful information of photos that, associated with visual features, could help create a more robust application for clothing annotation. For the realistic scenario, we created a vocabulary containing relevant terms related to different garments items using the tags and comments crawled with the images.

After filtering out all terms not in the vocabulary, the remaining textual content is described with TF-IDF vectors, which presents successful results in multi-modal approaches in the literature [14]. The TF-IDF transformation weights each term according to its discriminative capacity. Textual similarity between two images is assessed using the standard cosine and BM25 measures [5]. It is important to highlight that textual features are pre-processed separately and independently of the visual ones. After the pre-processing, these features are combined and used as input for the learning algorithm.

5.4 Baseline

The M3LDA algorithm [35] was used as baseline in this work. This method is a representative of the state-of-the-art in multi-label, multi-modal and multi-instance image annotation. It uses Latent Dirichlet Allocation (LDA) to create a rank with the most likely labels of each test image. This method provides superior mean Average Precision (mAP) numbers when compared against popular algorithms such as two MIML models (RankLoss [10] and DBA [73]) and two annotation models that allow region annotation (TM [13] and Corr-LDA [8]).

5.5 Cross Validation

For both scenarios, we conducted k-fold cross-validation in order to evaluate the algorithms. According to this protocol, a dataset is randomly split into k mutually exclusive subset (folds) of almost the same size. For the ideal scenario, the $k-1$ subsets are chosen as training set, and the remaining one is the test set. To work with all the dataset, the cross-validation process is repeated k times, and each time a subset is chosen to be the test set (without repetition). For the realistic scenario, $k-2$ subsets are chosen as training set, one fold is used as test-set, and the remaining one is the validation-set (in order to build the MDL function, as presented in 4.1.1). This last set is only used in the latter scenario, because in the former one we predict only one class per image, i.e., there is no need to build the MDL function. The process is repeated k times, and each time a subset is chosen to be the validation set while other subset is chosen to be the test one (without repetition), working with all dataset. At the end, the cross-validation estimate the arithmetic mean of all runs and the

Table 3: Cross-validation in different scenarios.

	Ideal Scenario (single-class)	**Realistic Scenario (multi-label)**
MMCA	Yes	Yes (MDL)
M3CA	No	Yes (MDL)

standard deviation between each one. The reported result is the average of the five runs.

Table 3 presents the cross-validation strategy used in both scenarios and proposed methods. For the ideal scenario, we used cross-validation without the validation-set when working with the MMCA method, since this scenario is composed of single-class classification. No experiments were conducted combining the ideal scenario and the M3CA approach. For the realistic scenario, we used cross-validation with the validation-set when working with both approaches, since this is a multi-label scenario and then the MDL function need to be built in order to define which labels should be assigned to the query image.

5.6 Evaluation Measures

To evaluate the experiments in the ideal scenario, the overall accuracy was used. For the realistic scenario, which may be categorized into a multi-label classification, we used the Jaccard distance as evaluation measure. Specifically, given the correct set of labels $\mathcal{L}_q^*$ and the predicted set of labels $\mathcal{L}_q$ for each query image q in the test-set $\mathcal{T}$, the Jaccard distance J is given as shown in Equation 5. It is important to highlight that Jaccard distance is considered as accuracy in multi-label tasks [59]. Moveover, this metric is the most similar to human perception as it considers all true positives, true negatives and false positives. Nowadays, it is a standard measure to evaluate image segmentation and annotation [15, 19, 12].

$$J = \frac{\sum \frac{|\{\mathcal{L}_q^* \cap \mathcal{L}_q\}|}{|\{\mathcal{L}_q^* \cup \mathcal{L}_q\}|}}{\mathrm{N}_q} \tag{5}$$

where N_q is the number of distinct query images in $\mathcal{T}$.

6 Results and Discussion

In this section, we present the experimental results to evaluate: (i) visual features, and (ii) proposed methods. When evaluating the visual features, we build the experiments in order to investigate how clothing annotation is impacted by different types of visual features. The second set of experiments, to evaluate the proposed methods, were devised to investigate: (i) the most

suitable approaches for the clothing annotation task, (ii) how each method is impacted by the proportion of relevant pixels, and (iii) how the proposed algorithms perform relatively to the baseline.

For investigating the presented items, we tested and varied some parameters to achieve more robust results. Concerning global descriptors, we have considered only the size of the association rule used on the classifier as a parameter. For BoVW [50], the observed parameters were the size of the rule (explained in Section 3) as well as the size of the visual dictionary $\mathcal{K}$ (the number of keywords generated by the mid-level representation). Regarding BossaNova approach [4], in addition to the parameters evaluated for the BoVW, we have observed the number of bins β used in the quantization step to encode the distances from one local descriptor to clusters. The default values for the size of the dictionary and the number of bins β were selected using a parameter evaluation made by [4].

In Section 6.1, we present the experimental evaluation of the visual feature descriptors followed by the evaluation of the proposed methods and baseline, presented in Section 6.2. For each evaluation process, we computed the mean processing time[5], in seconds, and the standard deviation based on five executions of each procedure.

6.1 Visual Features Evaluation

In this section, we present the experimental results carried out for evaluating the visual descriptors. As introduced in Section 5, we use the overall accuracy and the ideal scenario for these experiments.

Figure 11 shows the overall accuracy for the global descriptors followed by the mean processing time, in seconds, for each descriptor. Each plot groups descriptors of the same type: color, texture and shape. Between the global descriptors presented in Section 3, the best ones yield overall accuracy around 25%, which includes BIC, CCV, GCH and LCH descriptors. ACC, EOAC, and LAS achieved lower accuracy (around 15%) and are good candidates to be discarded on our next experiments.

Figure 12 shows the overall accuracy for the BoVW using different types of local descriptors and the mean processing time of each one. The plot represents the results varying the size of the visual dictionary (or feature vector) $\mathcal{K}$, which was chosen based on a parameter study made by [4]. Through the plot, it is possible to see that SIFT descriptors yields a good accuracy with any configuration of $\mathcal{K}$. It is also possible to observe that when $\mathcal{K} = 1024$ the proposed approach spends much less time if compare with the others.

Figure 13 shows the overall accuracy for the BoVW using SIFT descriptor and its processing time. The plot represents the results varying the size of the visual dictionary (or histogram of a image) $\mathcal{K}$. According to the plot, one can see that $\mathcal{K} = 1024$ yields a good accuracy (27%) if compared with the others.

[5] The processing time computed is only the time spent by the classification algorithm.

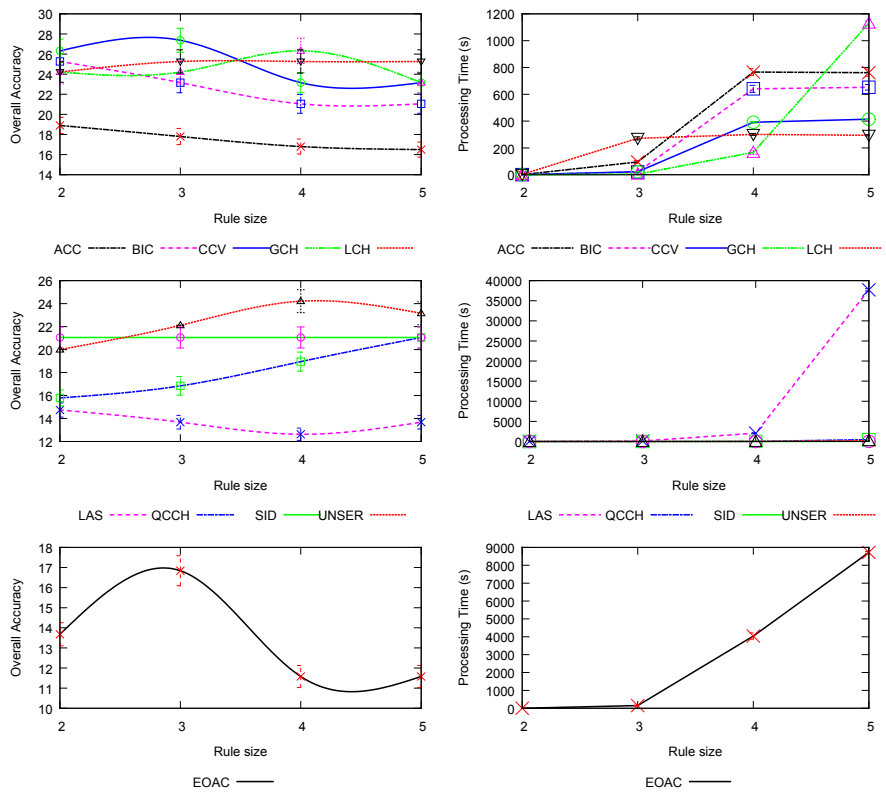


Fig. 11: The overall accuracy (left) and the processing time (right), in seconds, obtained using global descriptors. First row shows accuracy numbers for color descriptors. Second one shows the accuracy for texture descriptors. Last row shows accuracy numbers for shape descriptor.

It is also the most stable, with is virtually the same accuracy for all values of rule size. However, the best results, around 36%, were achieved with $\mathcal{K} = 2048$ and rule size 2, but with the increasing of the rule size, the accuracy tends to decrease.

Figure 14 shows the overall accuracy of BossaNova using different types of local descriptors, followed by the mean processing time of each one. The plot represents the results varying the size of the visual dictionary (or histogram of a image) $\mathcal{K}$, which was chosen based on a parameter study made by [4], and preserving the number of bins used in the quantization step in $\beta = 2$.

Note that, according to these results, SIFT descriptor is the most consistent one, since it yields good results practically independent of the configuration of $\mathcal{K}$. ORB descriptor, for example, yields good results when $\mathcal{K} = 1024$ and $\mathcal{K} = 4096$, but lower accuracy when $\mathcal{K} = 2048$. Another example is the SURF descriptor, that yields good results when $\mathcal{K} = 2048$ and $\mathcal{K} = 4096$, but not so

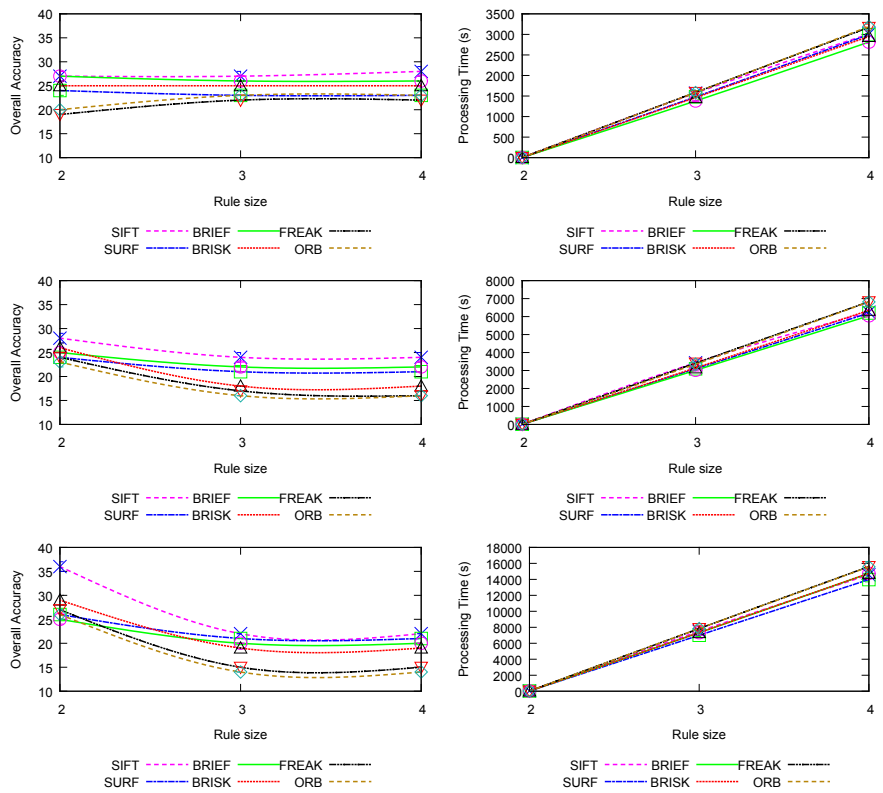


Fig. 12: The overall accuracy (left) and the processing time (right), in seconds, obtained with BoVW using hard assignment coding and max pooling in different local descriptors. First row shows the results with $\mathcal{K} = 1024$. Second one shows the results with $\mathcal{K} = 2048$, and the third shows the overall accuracy with $\mathcal{K} = 4096$.

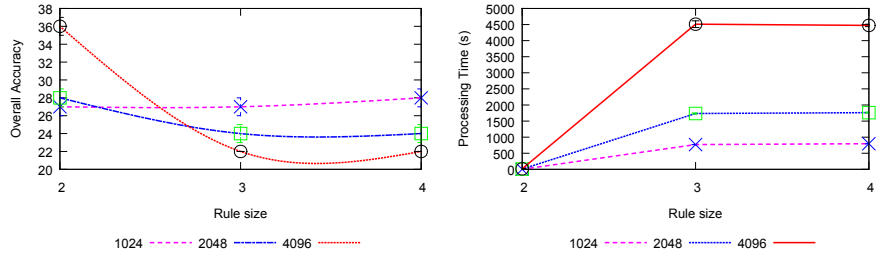


Fig. 13: The overall accuracy (left) and the processing time (right), in seconds, obtained with BoVW+SIFT, using hard assignment coding and max pooling.

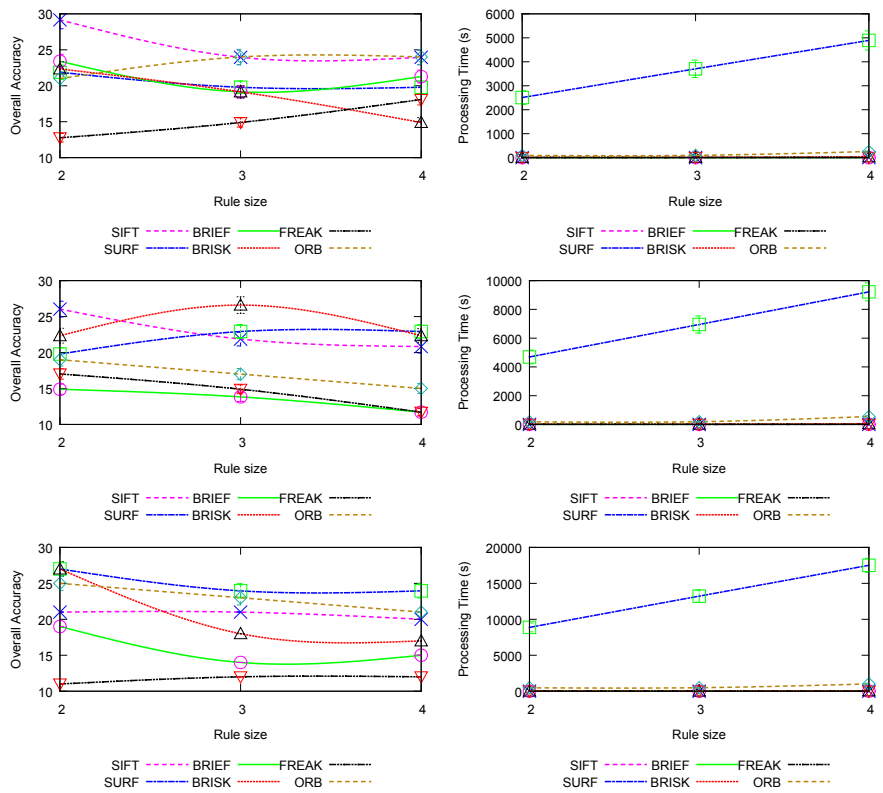


Fig. 14: The overall accuracy and the processing time, in seconds, obtained with BossaNova using different local descriptors. First row shows the results with $\mathcal{K} = 1024$. Second one shows the results with $\mathcal{K} = 2048$, and the third shows the overall accuracy with $\mathcal{K} = 4096$. For all results, the numbers of bins β used in the quantization step was fixed in 2.

good with $\mathcal{K} = 1024$. It is also possible to observe that when $\mathcal{K} = 1024$ the processing time spends much less time if compare with the others.

Figure 15 shows the accuracy for the BossaNova approach using only SIFT descriptor, with different dictionary sizes $\mathcal{K}$ and the numbers of bins β used in the quantization step. The values were defined based on a parameter evaluation study conducted by [4]. For the parameter β, three different values were evaluated: 2, 3 and 4. However, the results were very similar for all these values. This happens due to a normalization made by the BossaNova approach while creating the histogram, since with the increase β the numbers of codewords with high value tends to decrease and the normalization tries to maintain only the codewords with higher value. Thus, we report only the results for $\beta = 2$. Through the plots, it is possible to see that $\mathcal{K} = 1024$ yields the best results.

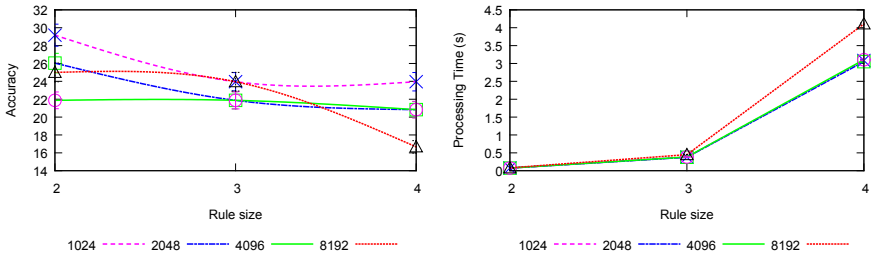


Fig. 15: The overall accuracy (left) and processing time (right) obtained for BossaNova using SIFT descriptor. The experiments were conducted using number of bins β with three different values: 2, 3 and 4. However, the results were very similar for all the values.

6.2 Proposed Methods Evaluation

For the experiments in realistic scenario, we have selected seven descriptors based on the experimental results presented in Section 6.1. The selected ones are those that yield at least 21% of accuracy: BIC, CCV, GCH, LCH, QCCH, SID and UNSER. For BoVW and BossaNova approaches, we chose only the best local descriptor: the SIFT one. In addition to this, we choose to create visual dictionaries with $\mathcal{K} = 1024$ since they achieve best and stable results.

It is also important to emphasize that based on the results in the ideal scenario, we could observe that smaller the size rule smaller the processing time and, in most cases, the accuracy tends to be very close for all variations of size rule. This allows us to conclude that the lowest value of size rule tends to be the best choice, since we capture the best benefit, i.e., we achieve good results in less processing time if compare with others size rules. Thus, all experiments in this section were made using size rule 2.

The results of the MMCA approach using each one of the visual descriptors are presented in Section 6.2.1. Section 6.2.2 presents the results of combinations of the MMCA method. Finally, a comparison between the proposed methods and the baseline are presented in Section 6.2.3.

6.2.1 MMCA Evaluation with Different Visual Descriptors

Figure 16 shows all the results for the evaluation of the methods. For each realistic scenario, we ran all feature descriptors using the MMCA approach and the results are shown in terms of Jaccard distance and standard deviation between the folds. It is possible to observe that, for most cases, SID descriptor is the best one amongst all of them. The BossaNova (BN) approach (using SIFT descriptor) is in second place in some cases, however, in general, mid-level approaches were not so effective, differing from the results observed from in the ideal scenario, where mid-level approaches were better than the global descriptors. This can be explained by the fact that, how the keypoints of the

mid-level strategies were extracted using dense sampling, when using a perfect segmentation mask, as in the ideal scenario, there is no effect of the background pixels in the codewords. However, if the mask do not perfectly adjust, wrong codewords may be created, interfering in the final result.

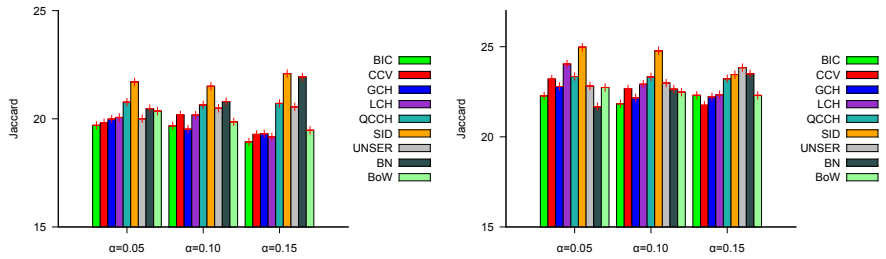


Fig. 16: Results for the MMCA method for Chictopia (left) and Pose (right).

6.2.2 Visual Descriptors Combination with MMCA

Figure 17 shows a comparison between the different combinations of the MMCA approach. In addition to the two combinations proposed in Section 4.1.2, we use two traditional ones: Condorcet Method (CM) and Borda Counting (BC). All combinations were evaluated using top-5, top-6 and top-7 approach. As expected, the combination of MMCA results yields better accuracy than the MMCA approach. The results were very similar, with Borda Counting being better, in most cases, for the Chictopia dataset, and Majority Probability being better, in most cases, for the Pose dataset.

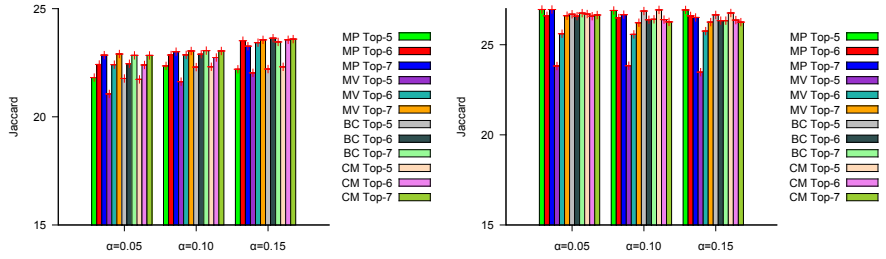


Fig. 17: The results of the combination of outputs of the MMCA for Chictopia (left) and Pose (right). Four combinations methods were compared: Majority Voting (MV), Majority Probability (MP), Condorcet Method (CM) and Borda Counting (BC).

6.2.3 Comparison with the Proposed Methods with the Baseline

Figure 18 shows a comparison between the proposed M3CA and the baseline (M3LDA). We also included the best results yielded using MMCA as well as using combination algorithm. As expected, accuracy increases with the number of features available. For both dataset, M3CA provides accuracy improvements that vary from 20% (M3LDA top-3) to 30% (M3LDA top-7). Through the figure, it is also possible to see that with the increasing of the mask α, the accuracy tends to increase. This reveal that small mask discard important visual features that may be used by the learning algorithm.

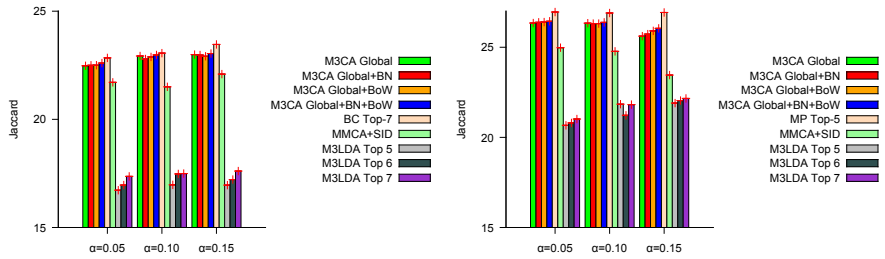


Fig. 18: The results of the M3CA and the baseline for Chictopia (left) and Pose (right). We also considered the best MMCA using SID, and the best combination algorithm for each dataset.

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.

Table 4 presents some examples of annotation of our proposed M3CA and the original annotations. First example shows a case when the algorithm could distinguish between several garment items but could not separate the sneakers from the ground, since both are very similar. Second example shows when the method could not distinguish the clothes, since all the garment items have the same color. Thus, the algorithm considered all the clothes (pants, shirt, sweater) as a single garment item, and predicted a romper. The last case is a perfect match of the predicted labels and the original ones. This case is only achieved when the function generated by the MDL suggested the perfect cut, predicting only good labels.

Table 4: Example of output of our proposed M3CA compared with the original annotations. In the third image, the predicted annotations are identical to the original ones.

Images			
Original Annotation	bag, hat, shorts, sneakers, sweater	bag, hat, heels, pants, shirt, sweater	bag, skirt, shoes, sweater
Automatic Annotation	shorts (0.12), sweater (0,09), shoes (0,08)	hat (0.10), romper (0.09)	skirt (0.11), bag (0.09), shoes, (0.08) sweater (0.08)

7 Final Remarks and Future Work

This paper presented a pointwise and a pairwise approach for clothing annotation. The first one, called MMCA, takes advantages from the multi-modal method resulting in a robust multi-label classification. The latter one, called M3CA, takes advantage of being multi-instance/multi-modal resulting also in a multi-label classification. It also exploits the benefits from different types of visual features. We also proposed two methods of combination of the pointwise results.

The novelty of this work relies on a multi-instance method that is capable of use the information from an image creating a sparse classifier. The performed experiments demonstrate the benefit of it, since it yields good results with less processing time when compared with state-of-the-art algorithm.

Considering the descriptors, SID descriptor is the best one amongst all of them. The BossaNova approach (using SIFT descriptor) is in second place in some cases, however, in general, mid-level approaches were not so effective for the realistic scenario, differing from the results observed from in the ideal scenario. This can be explained by the fact that, if the mask do not perfectly adjust, wrong codewords may be created, interfering in the final result of the mid-level strategies. Because SID descriptors analyze the image as a whole, this problem is softened.

In three cases, the best result for the investigated problem were achieved using combinations for the output of the MMCA approach: Majority Probability yields best results for Pose dataset in two cases, while Majority Voting achieves best results for Chictopia in one configuration. For the remain cases, Borda Counting achieves the best results in the last two configurations of the Chictopia dataset while Condorcet Method achieves the best results in the last

case for the Pose Dataset. However, in the cases where the proposed combination methods loses, it generally stays close to the best results, which makes this an advantage, since they are easier to implement than the traditional ones.

Though, the combination of MMCA results yields better accuracy than the M3CA approach, the MMCA approach, without combination, was not capable to achieve accuracy similar 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 spent, since to get the combination, we need to get all results from each descriptor.

Although M3CA is designed for clothing annotation, it is possible to be applied to others tasks. In the future, we plan to adapt the proposed M3CA for clothing parsing and also using M3CA with different learning techniques.

Acknowledgements The authors would like to acknowledge grants from CNPq (grant 449638/2014-6), CAPES, Fundação de Apoio à Pesquisa do Estado de Minas Gerais (Fapemig, under the grant APQ-00768-14), PRPq/Universidade Federal de Minas Gerais, Finep, and InWeb − the Brazilian National Institute of Science and Technology for the Web.

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: International Conference on Management of Data, pp 207–216
2. Alahi A, Ortiz R, Vandergheynst P (2012) FREAK: fast retina keypoint. In: Conference on Computer Vision and Pattern Recognition, pp 510–517
3. Atrey PK, Hossain MA, El-Saddik A, Kankanhalli MS (2010) Multimodal fusion for multimedia analysis: a survey. Multimedia System 16(6):345–379
4. de Avila SEF, Thome N, Cord M, Valle E, de Albuquerque Araújo A (2011) BOSSA: extended bow formalism for image classification. In: International Conference on Image Processing, pp 2909–2912
5. Baeza-Yates RA, Ribeiro-Neto BA (2011) Modern Information Retrieval - the concepts and technology behind search, 2nd edn. Pearson Education Ltd., Harlow, England
6. Bay H, Ess A, Tuytelaars T, Gool LJV (2008) Speeded-up robust features (SURF). Computer Vision and Image Understanding 110(3):346–359
7. Bekele D, Teutsch M, Schuchert T (2013) Evaluation of binary keypoint descriptors. In: International Conference on Image Processing, pp 3652–3656
8. Blei DM, Jordan MI (2003) Modeling annotated data. In: ACM Special Interest Group on Information Retrieval, pp 127–134
9. Boureau Y, Bach F, LeCun Y, Ponce J (2010) Learning mid-level features for recognition. In: Conference on Computer Vision and Pattern Recognition, pp 2559–2566
10. Briggs F, Fern XZ, Raich R (2012) Rank-loss support instance machines for miml instance annotation. In: International Conference on Knowledge Discovery and Data Mining, pp 534–542

11. Calonder M, Lepetit V, Strecha C, Fua P (2010) BRIEF: binary robust independent elementary features. In: European Conference on Computer Vision, pp 778–792
12. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp 248–255
13. Duygulu P, Barnard K, de Freitas JFG, Forsyth DA (2002) Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: European Conference on Computer Vision, pp 97–112
14. Escalante HJ, Montes M, Sucar E (2012) Multimodal indexing based on semantic cohesion for image retrieval. Information Retrieval 15(1):1–32
15. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2):303–338
16. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: International Joint Conference on Artificial Intelligence, pp 1022–1029
17. Feng S, Xu D (2010) Transductive multi-instance multi-label learning algorithm with application to automatic image annotation. Expert Systems with Applications 37(1):661–670
18. Gallagher AC, Chen T (2008) Clothing cosegmentation for recognizing people. In: Conference on Computer Vision and Pattern Recognition
19. Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: The kitti dataset. International Journal of Robotics Research
20. van Gemert J, Geusebroek J, Veenman CJ, Smeulders AWM (2008) Kernel codebooks for scene categorization. In: European Conference on Computer Vision, pp 696–709
21. Guillaumin M, Mensink T, Verbeek JJ, Schmid C (2009) Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: International Conference on Computer Vision, pp 309–316
22. Guillaumin M, Verbeek JJ, Schmid C (2010) Multimodal semi-supervised learning for image classification. In: Conference on Computer Vision and Pattern Recognition, pp 902–909
23. Huang C, Liu Q (2007) An orientation independent texture descriptor for image retireval. In: International Conference on Computer and Computational Sciences, pp 772–776
24. Huang J, Kumar R, Mitra M, Zhu W, Zabih R (1997) Image indexing using color correlograms. In: Conference on Computer Vision and Pattern Recognition, pp 762–768
25. Kalantidis Y, Kennedy L, Li L (2013) Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In: International Conference on Multimedia Retrieval, pp 105–112
26. Leutenegger S, Chli M, Siegwart R (2011) BRISK: binary robust invariant scalable keypoints. In: International Conference on Computer Vision, pp 2548–2555

27. Li R, Lu J, Zhang Y, Zhao T (2010) Dynamic adaboost learning with feature selection based on parallel genetic algorithm for image annotation. Knowledge-Based Systems 23(3):195–201
28. Liu S, Song Z, Liu G, Xu C, Lu H, Yan S (2012) Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In: Conference on Computer Vision and Pattern Recognition, pp 3330–3337
29. Liu T (2009) Learning to rank for information retrieval. Foundations and Trends in Information Retrieval 3(3):225–331
30. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110
31. Mahmoudi F, Shanbehzadeh J, Eftekhari-Moghadam A, Soltanian-Zadeh H (2003) Image retrieval based on shape similarity by edge orientation autocorrelogram. Pattern Recognition 36(8):1725–1736
32. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation. In: European Conference on Computer Vision, Springer-Verlag, pp 316–329
33. Maron O, Lozano-Pérez T (1997) A framework for multiple-instance learning. In: Neural Information Processing Systems, pp 570–576
34. Moran S, Lavrenko V (2014) Sparse kernel learning for image annotation. In: International Conference on Multimedia Retrieval, p 113
35. Nguyen C, Zhan D, Zhou Z (2013) Multi-modal image annotation with multi-instance multi-label LDA. In: International Joint Conference on Artificial Intelligence
36. Nogueira K, Veloso AA, dos Santos JA (2014) 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, pp 327–334
37. Ntalianis K, Tsapatsoulis N, Doulamis A, Matsatsinis N (2014) Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution. Multimedia Tools and Applications 69(2):397–421
38. Oliva A, Torralba A (2006) Building the gist of a scene: The role of global image features in recognition. Visual Perception, Progress in Brain Research 155:23–36
39. Pass G, Zabih R, Miller J (1996) Comparing images using color coherence vectors. In: International Conference on Multimedia, pp 65–73
40. Penatti OAB, Valle E, da Silva Torres R (2012) Comparative study of global color and texture descriptors for web image retrieval. Journal of Visual Communication and Image Representation 23(2):359–380
41. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2008) Lost in quantization: Improving particular object retrieval in large scale image databases. In: Conference on Computer Vision and Pattern Recognition
42. Read J, Pfahringer B, Holmes G (2008) Multi-label classification using ensembles of pruned sets. In: International Conference on Data Mining, pp 995–1000

43. Rublee E, Rabaud V, Konolige K, Bradski GR (2011) ORB: an efficient alternative to SIFT or SURF. In: International Conference on Computer Vision, pp 2564–2571
44. dos Santos JA, Penatti OAB, da Silva Torres R (2010) Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. In: International Conference on Computer Vision Theory and Applications, pp 203–208
45. dos Santos JA, Faria FA, da Silva Torres R, Rocha A, Gosselin PH, Philipp-Foliguet S, Falcão AX (2012) Descriptor correlation analysis for remote sensing image multi-scale classification. In: International Conference on Pattern Recognition, pp 3078–3081
46. Shen EY, Lieberman H, Lam F (2007) What am I gonna wear?: scenario-oriented recommendation. In: International Conference on Intelligent User Interfaces, pp 365–368
47. da Silva Torres R, Falcão AX (2006) Content-based image retrieval: Theory and applications. RITA 13(2):161–185
48. Simo-Serra E, Fidler S, Moreno-Noguer F, Urtasun R (2014) A High Performance CRF Model for Clothes Parsing. In: Asian Conference on Computer Vision
49. Simo-Serra E, Fidler S, Moreno-Noguer F, Urtasun R (2015) Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. In: Conference on Computer Vision and Pattern Recognition
50. Sivic J, Zisserman A (2006) Video google: Efficient visual search of videos. In: Toward Category-Level Object Recognition, pp 127–144
51. Snow R, O'Connor B, Jurafsky D, Ng AY (2008) Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Conference on Empirical Methods in Natural Language Processing, pp 254–263
52. Socher R, Lin CC, Ng AY, Manning CD (2011) Parsing natural scenes and natural language with recursive neural networks. In: International Conference on Machine Learning, pp 129–136
53. Stehling RO, Nascimento MA, Falcão AX (2002) A compact and efficient image retrieval approach based on border/interior pixel classification. In: International Conference on Information and Knowledge Management, pp 102–109
54. Suh B, Bederson BB (2007) Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition. Interacting with Computers 19(4):524–544
55. Swain MJ, Ballard DH (1991) Color indexing. International Journal of Computer Vision 7(1):11–32
56. Tang J, Li H, Qi G, Chua T (2010) Image annotation by graph-based inference with integrated multiple/single instance representations. IEEE Transactions on Multimedia 12(2):131–141
57. Tao B, Dickinson BW (2000) Texture recognition and image retrieval using gradient indexing. Journal of Visual Communication and Image Representation 11(3):327–342

58. Tokumaru M, Fujibayashi T, Muranaka N, Imanishi S (2002) Virtual stylist project - dress up support system considering user's subjectivity. In: International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age, pp 207–211
59. Tsoumakas G, Katakis I (2006) Multi-label classification: An overview. Dept of Informatics, Aristotle University of Thessaloniki, Greece
60. Tsoumakas G, Katakis I (2007) Multi-label classification: An overview. International Journal of Data Warehouse and Mining 3(3):1–13
61. Tuytelaars T (2010) Dense interest points. In: Conference on Computer Vision and Pattern Recognition, pp 2281–2288
62. Tuytelaars T, Mikolajczyk K (2007) Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3):177–280
63. Unser M (1986) Sum and difference histograms for texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(1):118–125
64. Veloso A, Jr WM, Zaki MJ (2006) Lazy associative classification. In: International Conference on Data Mining, pp 645–654
65. Veloso A, Jr WM, Gonçalves MA, Zaki MJ (2007) Multi-label lazy associative classification. In: Conference on Principles and Practice of Knowledge Discovery in Databases, pp 605–612
66. Vens C, Struyf J, Schietgat L, Dzeroski S, Blockeel H (2008) Decision trees for hierarchical multi-label classification. Machine Learning 73(2):185–214
67. Vogiatzis D, Pierrakos D, Paliouras G, Jenkyn-Jones S, Possen BJHHA (2012) Expert and community based style advice. Expert Systems with Applications 39(12):10,647–10,655
68. Weber M, Bäuml M, Stiefelhagen R (2011) Part-based clothing segmentation for person retrieval. In: International Conference on Advanced Video and Signal-Based Surveillance, pp 361–366
69. Xie L, Pan P, Lu Y (2015) Markov random field based fusion for supervised and semi-supervised multi-modal image classification. Multimedia Tools and Applications pp 613–634
70. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2012) Parsing clothing in fashion photographs. In: Conference on Computer Vision and Pattern Recognition, pp 3570–3577
71. Yamaguchi K, Kiapour MH, Berg TL (2013) Paper doll parsing: Retrieving similar styles to parse clothing items. In: International Conference on Computer Vision, pp 3519–3526
72. Yang M, Yu K (2011) Real-time clothing recognition in surveillance videos. In: International Conference on Image Processing, pp 2937–2940
73. Yang S, Zha H, Hu B (2009) Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In: Neural Information Processing Systems, pp 2143–2150
74. Yang Y, Ramanan D (2011) Articulated pose estimation with flexible mixtures-of-parts. In: Conference on Computer Vision and Pattern Recognition, pp 1385–1392

75. Zegarra J, Leite N, Torres R (2008) Wavelet-based feature extraction for fingerprint image retrieval. Journal of Computational and Applied Mathematics
76. Zhang D, Lu G (2004) Review of shape representation and description techniques. Pattern Recognition 37(1):1–19
77. Zhang D, Islam MM, Lu G (2012) A review on automatic image annotation techniques. Pattern Recognition 45(1):346–362
78. Zhaolao L, Zhou M, Wang X, Fu Y, Tan X (2013) Semantic annotation method of clothing image. In: International Conference on Human-Computer Interaction, pp 289–298
79. Zhou Z, Zhang M, Huang S, Li Y (2012) Multi-instance multi-label learning. Artificial Intelligence 176(1):2291–2320