Learning to Annotate Clothes in Everyday Photos: Multi-Modal, Multi-Label, Multi-Instance Approach

Keiller Nogueira, Adriano Veloso, and Jefersson A. dos Santos
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Email: {keillernogueira, adrianov, jefersson}@ufmg.br

***Abstract*—In this paper, we present an effective algorithm to automatically annotate clothes in everyday photos posted in online social networks, such as Facebook and Instagram. Specifically, clothing annotation can be informally stated as predicting, as accurately as possible, the garment items appearing in the target photo. This task not only poses interesting challenges for existing vision and recognition algorithms, but also brings huge opportunities for recommender and e-commerce systems. We formulate the annotation task as a multi-modal, multi-label and multi-instance classification problem: (i) both image and textual content (i.e., comments about the image) are available for learning classifiers, (ii) the classifiers must predict a set of labels (i.e., a set of garment items), and (iii) the decision on which labels to predict comes from a *bag* of instances that are used to build a function, which separates labels that should be predicted from those that should not be. Under this setting, we propose a classification algorithm which employs association rules in order to build a prediction model that combines image and textual information, and adopts an entropy-minimization strategy in order to the find the best set of labels to predict. We conducted a systematic evaluation of the proposed algorithm using everyday photos collected from two major fashion-related social networks, namely pose.com and chictopia.com. Our results show that the proposed algorithm provides improvements when compared to popular first choice multi-label algorithms that range from 2% to 40% in terms of accuracy.**

***Keywords*-Automatic annotation; multi-label classification; machine learning;**

I. Introduction

Online Social Networks allow their members to express themselves in different ways, by creating and sharing content. A particular way of expression being increasingly adopted is to post photos showing their latest looks and clothes. Typically, shortly after the image is shared, other members also participate by posting comments concerning the garment items in the image. This results in torrents of clothing-related image-text content, and also on huge opportunities for recommender and e-commerce systems. In this paper we are particularly interested in the clothing annotation task, which can be stated as the task of assigning short textual descriptors or keywords (called tags) to images. Such tags are related to specific garment items, such as shirts, trousers and shoes, and multiple tags may be associated with an arbitrary image.

A. Our Approach to Clothing Annotation

We formulate the clothing annotation task as a 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 predicts the labels/tags that are more likely to be associated with it. Next we discuss the basic components of our solution.

a) Features: In our setting, we model each data instance as a pair of images: the target image q and the base image b. Labels associated with base image b are always known in advance (i.e., base labels), but we want to predict labels for the target image q. Features that compose each instance (q, b) are given as the set of labels associated with image b, plus a set of distances between images q and b. Such distances are computed using image content descriptors [1], [2], [3], [4], as well as terms within comments surrounding both images. The intuition is that similar images are likely to share common labels, and thus small distances are expected to increase the membership probabilities associated with the correct labels for the target image q.

b) Classification Model: Our classifiers are composed of association rules [5], which are essentially local mappings $X \rightarrow y$ relating a combination of distances+base labels 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 images in the test-set, our algorithm builds multiple small classifiers, one for each image in the test-set.

c) Prediction: Typical solutions to multi-label classification employ the top-k approach, where a pre-determined threshold k is used to select the labels to predict. That is, only the k labels with the highest membership probabilities are predicted. Instead of relying on this parameter, we propose an entropy-minimization multi-instance approach which finds a different cut point for each instance (q, b) in the test-set. More specifically, we build a function $\gamma(\mathcal{L}_{(q,b)})$, which receives as input a set of candidate labels $\mathcal{L}_{(q,b)}$ for instance (q, b), and returns the best cut for these labels. Finally, only labels with membership probability exceeding the cut point are predicted.

B. Contributions and Findings

In practice, we claim the following benefits and contributions over existing solutions:

- Our main contribution is a novel algorithm to automatic clothing annotation. Our proposed algorithm learns effective classifiers and decides the best set of labels to predict by following a multi-modal, multi-label, multi-instance approach.
- 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 [6], by a factor that ranges from 2% to 40% in terms of standard accuracy measures.

The paper is structured as follows. Related work is presented in Section 2. We introduce basic concepts as well as our clothing annotation algorithm in Section 3. Experimental evaluation, as well as the effectiveness of the proposed algorithm, is discussed in Section 4. Finally, in Section 5 we conclude the paper and point promising directions for future work.

II. Related Work

Automatic image annotation approaches that use supervised machine learning algorithms are becoming increasingly popular [7]. Firstly, main approaches towards image annotation modelled the learning problem as machine translation [8] or correlation learning tasks [9], [10]. Some other researchers regarded the automatic image annotation task as multi-label classification problem [11], [12]. Multi-label classification algorithms can be categorized into two different groups [13]:

1) problem transformation methods and
2) algorithm adaptation methods.

The first group includes methods that are algorithm independent, that is, they transform the multi-label problem into one or more single-label problems. There is a lot of work on this group that includes binary relevance method [14], [15], binary pairwise classification approach [16] and label combination or label power-set method (LC) [17]. The second group includes methods that extend specific learning algorithms in order to handle multi-label data directly. Well-known approaches include Adaboost [18], decision trees [19], lazy methods [6], [20], [21], [22], [23], and neural networks [24].

From the view point of multi-modal learning [25] our proposed algorithm follows the feature-level fusion approach, where different types of features are extracted from input data and then combined and sent to the learning algorithm. From the view point of the fashion application scenario, there are many works related to clothing recognition [23], [26], [27]. Next, we introduce the most relevant works related to ours.

Tokumaru et al. [28] proposed a system, named "Virtual Stylist", which aims to help users to find out outfits that might fit them well. The work of [29] describes the recommendation of clothes based on the similarity between users and models appearing in fashion magazines. In [23], the authors worked on the clothing parsing problem using a retrieval based approach. For a query image, they find similar styles from a large database of tagged fashion images and use these examples to parse the query. Focusing on works that exploit the relation between visual and textual, we can point out [30] which introduces the recommendation of outfits for specific occasions based on textual input that defines the occasion and how the user wants to look like.

III. Multi-Modal, Multi-Label, Multi-Instance Algorithm for Clothing Annotation

In this section we present our algorithm for automatic clothing annotation. Our algorithm, named Multi-Modal/Multi-Label/Multi-Instance Clothing Annotation algorithm (or simply M3CA), builds classifiers on a demand-driven basis. Each classifier returns membership probabilities for each label, and the final set of labels to predict comes by minimizing the entropy of such membership probabilities. We start discussing visual and textual elements used to represent our instances.

A. Visual and Textual Descriptors

Images posted in online social networks (in particular those related to clothing) may contain both visual and textual elements, and each modality may be analyzed in a variety of ways. For instance, visual elements can be analyzed based on color, texture, shape, and so on. Textual elements, in turn, may include terms related to garment items. Specifically, we observed that:

1) Images sharing common garment items are likely to share similar visual elements (e.g., color, texture and shape), and,
2) People tend to post similar terms in comments associated with images that share common garment items.

Image Descriptors. Visual elements are strongly based upon the concept of image descriptors [31]. A descriptor expresses perceptual qualities of an image, and is composed by

1) A feature-vector that represents image properties, such as color, texture and shape, and,
2) A distance function that returns the similarity between two images as a function of the distances between their corresponding feature-vectors.

Both the feature-vector and the distance function affect how the descriptor encode the perceptual qualities of the images. An image descriptor representation to compute the distance between two input images is presented in Figure 1.

There is a multitude of descriptors available in the literature [32] that can used to represent visual elements. Clearly, different descriptors produce different results. Further, it is intuitive that different descriptors may provide complementary information about images, so that 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 and unlikely to obtain in advance. Specifically, we used a set of 11 descriptors based on color, texture and shape, in order to extract visual features from each image:

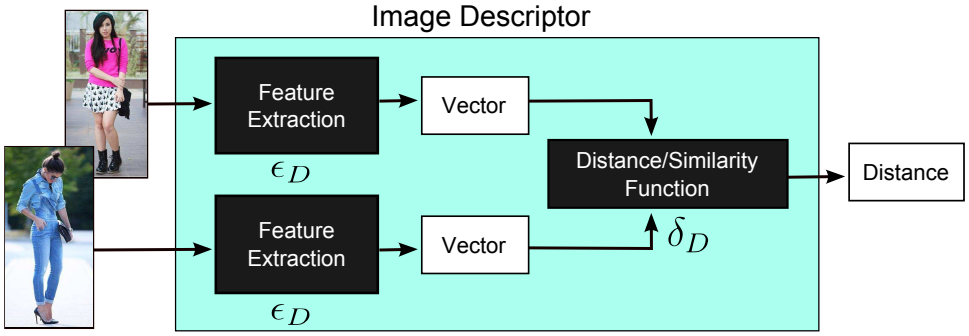

Fig. 1. An image descriptor representation.

1) Color descriptors:
   a) Auto-Correlogram Color (ACC) [33]
   b) Border/Interior Pixel Classification (BIC) [1]
   c) Color Coherence Vector (CCV) [34]
   d) Global Color Histogram (GCH) [35]
   e) Local Color Histogram (LCH) [35]
2) Texture descriptors:
   a) Quantized Compound Change Hist. (QCCH) [36]
   b) Local Activity Spectrum (LAS) [37]
   c) Steerable Pyramid Decomposition (SID) [4]
   d) Unser [2]
3) Shape descriptors:
   a) Edge orientation auto-correlogram (EOAC) [3]

The aforementioned descriptors are used to calculate distances between pairs of images. We used the common L1 distance to calculate the similarity between two images. It calculates the distance between two vectors by sum the modulus of the difference between each value of the vector. These distances are used as features for our classification algorithm.

Textual Information. Textual content includes terms appearing in comments associated with a pair of images. We define a vocabulary containing terms related to different garments items. After filtering out all terms not in the vocabulary, the remaining textual content is described with TF-IDF vectors. 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 [38].

B. Learning Classifiers

We provide a set of pairs of images as input to our algorithm. A pair of images is denoted as an instance.

Definition 1. An instance (q, b) is composed by a base image b and a target image q. Labels associated with the base image b are called *base labels* and are always known in advance. Labels associated with the target image q must be predicted. An instance (q, b) is represented by a set of (visual and textual) distances between q and b, along with the base labels. Specifically, an instance is represented as a list $(q, b) = \{f_1, f_2, \ldots, f_m, v_1, v_2, \ldots, v_n\}$ of m distance values and n base labels.

The M3CA Algorithm. Our proposed multi-modal, multi-label, multi-instance clothing annotation algorithm uses association rules [5] to produce classifiers that predict garment items associated with an arbitrary image. The M3CA algorithm receives as input a labeled training-set $\mathcal{D}$ composed of records of the form $< q, B >$, where q is a target image and B is a bag of base images. The bag B is partitioned into multiple instances of the form $(q, b, \mathcal{L}_q^*)$, where $b \in B$ and $\mathcal{L}_q^*$ is a set of labels associated with the target image q (i.e., the garment items appearing in image q). Distances in (q, b) are discretized [39] and then assigned to distance intervals,[1] in order to allow for the enumeration of association rules. Labels are drawn from a set of 31 discrete possibilities, including "trousers", "glasses", "shirts", "shoes", and "tennis".

The test-set $\mathcal{T}$ also consists of records of the form $< q, B >$. Again, the bag B is partitioned into multiple instances $(q, b, ?)$. 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. From the training-set $\mathcal{D}$, the algorithm extracts a rule-set $\mathcal{R}$ composed of garment rules which are used to predict a set of labels $\mathcal{L}_{(q,b)}$ which approximates as accurately as possible $\mathcal{L}_q^*$.

Definition 2. A garment rule has the form:

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

where $j \geq 1$ and $z \leq m$, and $t \geq 1$ and $u \leq n$.

The operator "and" represents that the antecedent of a rule is formed with the simple presence of a determined combination of features and base labels. These combinations work like a signature to the rule.

These rules can contain any mixture of distance intervals and base labels in the antecedent and a label l_i (i.e., a garment

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

item) in the consequent. The strength of the association between antecedent and consequent is measured by a statistic θ, which is known as confidence [5] and is simply the conditional probability of the consequent given the antecedent. Basically, each garment rule $\{X \rightarrow l_i\} \in \mathcal{R}$ is a vote given for label l_i. Given an instance $(q, b) \in \mathcal{T}$, a garment rule is a valid vote if it is applicable to (q, b).

Definition 3. A garment rule $\{X \rightarrow l_i\}$ is said to be applicable to instance $(q, b) \in \mathcal{T}$ if all intervals and base labels in X are in (q, b), that is, $X \subseteq (q, b)$.

We denote as $\mathcal{R}_{(q,b)}$ the set of garment rules in $\mathcal{R}$ that are applicable to instance (q, b). Thus, only rules in $\mathcal{R}_{(q,b)}$ are considered as valid votes when predicting the labels for image q. Further, we denote as $\mathcal{R}^{l_i}_{(q,b)}$ the subset of $\mathcal{R}_{(q,b)}$ containing only rules predicting label l_i. Votes in $\mathcal{R}^{l_i}_{(q,b)}$ have different weights, depending on the confidence of the corresponding rules. Given an arbitrary target image q and a base image $b \in B$, the weighted votes for label l_i are averaged, resulting in the score for l_i, as shown in Equation 1:

$$s(q, b, l_i) = \frac{\sum \theta(X \rightarrow l_i)}{|\mathcal{R}^{l_i}_{(q,b)}|} \text{ , where } X \subseteq (q, b). \quad (1)$$

where $|\mathcal{R}|$ represents the set size of the garment rules.

The likelihood of target image q being associated with label l_i is obtained by normalizing the scores, as expressed by $\hat{p}(l_i|(q, b))$, shown in Equation 2:

$$\hat{p}(l_i|(q, b)) = \frac{s(q, b, l_i)}{\sum_j s(q, b, l_j)}. \quad (2)$$

In this case, higher values of $\hat{p}(l_i|(q, b))$ indicate lower distances between images q and b, and labels associated with b are also likely to be associated with q. On the other hand, lower values of $\hat{p}(l_i|(q, b))$ indicate higher distances between images q and b, and thus labels associated with b are not likely to be associated with q.

Minimum-Entropy Cut. Given an instance (q, b) and a set of candidate labels $\mathcal{L}_{(q,b)}$ provided by the classifier,[2] we must find a cut point $c_{(q,b)}$ which delimits labels that are likely to be associated with target image q from those that are not. In other words, we must find a threshold $c_{(q,b)}$, so that only labels in $\mathcal{L}_{(q,b)}$ for which $\hat{p}(l_i|(q, b)) > c_{(q,b)}$ are finally predicted.

Our approach searches for a threshold $c_{(q,b)}$ that provides the best entropy cut in the space induced by probabilities $\hat{p}(l_i|(q, b))\ \forall\ l_i \in \mathcal{L}_{(q,b)}$. Figure 2 illustrates our approach. In the figure, symbol $^-$ indicates that the corresponding label l_i is associated with target image q. Similarly, symbol $^\circ$ indicates that the corresponding label l_i is not associated with target image q. Therefore, in the example, labels $\{l_4, l_5, l_6\}$ are associated with q (i.e., $^-$), while labels $\{l_1, l_2, l_3\}$ are not (i.e., $^\circ$). 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 our approach is general enough to handle such harder cases. The basic idea is that any value of $c_{(q,b)}$ induces two partitions over the space of values for $\hat{p}(l_i|(q, b))$, that is, one partition with probabilities that are lower than $c_{(q,b)}$, and another partition with probabilities higher than $c_{(q,b)}$. Our approach sets $c_{(q,b)}$ to the value that minimizes the average entropy of these two partitions.

Definition 4. Consider a list $\mathcal{O} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_j \in \{^\circ, ^-\}$ and y_j is a membership probability $\hat{p}(l_i|(q, b))$. The list is sorted such that $y_j \leq y_{j+1}$. Also consider c as a candidate value for $c_{(q,b)}$. 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.

Our approach works as follows. Firstly, it calculates the entropy in $\mathcal{O}$, as shown in Equation 3. Then, it calculates the sum of the entropies in each partition induced by c, according to Equation 4. Finally, it sets $c_{(q,b)}$ to the value of c that minimizes $E(\mathcal{O}) - E(\mathcal{O}_c)$.

$$\begin{aligned} E(\mathcal{O}) = & - \left(\frac{\mathrm{N}^\circ(\mathcal{O})}{|\mathcal{O}|} \times \log \frac{\mathrm{N}^\circ(\mathcal{O})}{|\mathcal{O}|} \right) \\ & - \left(\frac{\mathrm{N}^-(\mathcal{O})}{|\mathcal{O}|} \times \log \frac{\mathrm{N}^-(\mathcal{O})}{|\mathcal{O}|} \right) \end{aligned} \quad (3)$$

where N° gives the number of labels in $\mathcal{L}_{(q,b)}$ but not in $\mathcal{L}^*_q$, and N^- gives the number of labels in $\mathcal{L}_q$ and also in $\mathcal{L}^*_q$.

$$\begin{aligned} E(\mathcal{O}_c) = & \frac{|\mathcal{O}_c(\leq)|}{|\mathcal{O}|} \times E(\mathcal{O}_c(\leq)) + \\ & \frac{|\mathcal{O}_c(>)|}{|\mathcal{O}|} \times E(\mathcal{O}_c(>)) \end{aligned} \quad (4)$$

We employ a validation-set $\mathcal{V}$ composed of several instances (q, b), so that both the true labels $\mathcal{L}^*_q$ and the predicted labels $\mathcal{L}_{(q,b)}$ are known for all instances in the validation-set. Our goal is to build a function $\gamma(\mathcal{L}_{(q,b)})$ which receives as inputs a set of candidate labels $\mathcal{L}_{(q,b)}$ and returns the best entropy cut for these labels. Function $\gamma(\mathcal{L}_{(q,b)})$ thus, is simply given as the average of the best entropy cuts associated with instances $(q, b) \in \mathcal{V}$ having $\mathcal{L}_{(q,b)}$ as candidate labels.

C. Prediction

A target image q may appear within several (i.e., n) instances $(q, b_i) \in \mathcal{T}$. For each instance $(q, b_i) \in \mathcal{T}$ a specific set

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

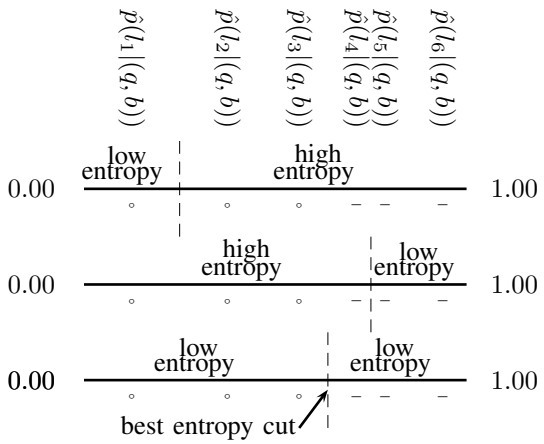

Fig. 2. Looking for the minimum entropy cut for a specific instance (q, b).

of labels $\mathcal{L}_{(q,b_i)}$ is associated with q. The final set of labels to be predicted is given as $\mathcal{L}_q = \{\mathcal{L}_{(q,b_1)} \cup \mathcal{L}_{(q,b_2)} \cup \ldots \cup \mathcal{L}_{(q,b_n)}\}$.

After aggregate the labels, we can select the ones with bigger probability and then, using the minimum-entropy cut, estimate the labels related with q.

IV. Experimental Evaluation

In this section, we present the experimental results for the evaluation of the proposed M3CA algorithm. We used the Jaccard distance as a standard accuracy measure. We consider the CLAC algorithm [6] in order to provide baseline comparison. The CLAC algorithm is a representative of the state-of-the-art in multi-label classification. It provides superior accuracy numbers when compared against popular algorithms such as BoosTexter [40] and MLSVM [41]. Next we present the datasets we used in our experiments, and then we discuss our results and main findings.

A. Datasets

We have crawled images, tags and associated comments from two fashion-related social networks, namely pose.com and chictopia.com. Basic information about the resulting datasets is shown in Table I. The pose.com dataset was crawled from January 15, 2014 to January 25, 2014. This resulted in more than three thousands images, but we discarded images with less than five associated comments. The chictopia.com dataset was crawled from January 25, 2014 to February 5, 2014. This resulted in more than two thousands images, but again, we discarded all images with less than five associated comments.

Figure 3 shows the frequency of each label. As expected, some labels occur frequently (e.g., "Bag", "Boots", and "Coat"), while others occur only few times (e.g., "Tie", "Stockings", and "Wallet"). Figure 4 shows the cumulative distribution function for labels in chictopia.com and pose.com. The probability for an arbitrary image having at least x labels decreases almost linearly in both cases.

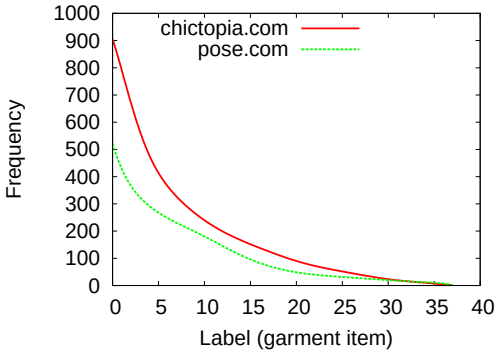

Fig. 3. Frequency distribution of labels.

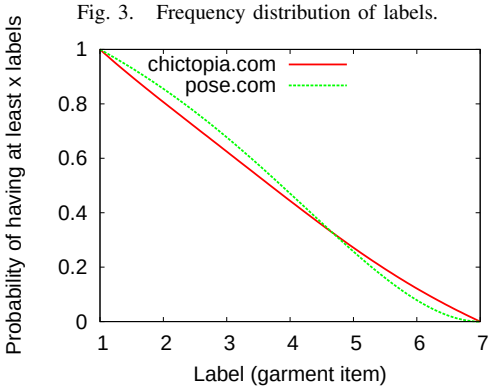

Fig. 4. Cumulative distribution function of labels.

TABLE I
Datasets.

	pose.com	chictopia.com
Number of photos	2,306	1,579
Number of tags	7,501	5,093
Number of comments	27,486	12,348
Tags per photo	3.25	3.23
Comments per photo	11.92	7.82

Image Segmentation. In order to avoid the effect of background pixels over the description of an image, we ran a human pose estimation algorithm [42] and separate relevant pixels. Pose estimation and image segmentation are needed because we are working with global image descriptors, and thus the background may input noise. Figure 5 shows the original image, a pose estimation skeleton [42] and relevant pixels. The final set of relevant pixels (i.e., non-background pixels) was obtained using the skeleton and employing a factor of proportionality in order to enlarge each line of the skeleton, thus encompassing the entire pose.

We discarded images according to the proportion of background/non-background pixels. More specifically, we discarded all images for which the proportion of relevant pixels (non-background pixels) is lower than a fixed threshold α_{min}. Table II shows the number of remaining images for different values of α_{min}.

Fig. 5. (Left) Original image. (Middle) Skeleton. (Right) Pixels of interest.

TABLE II
IMAGES WITH ENOUGH RELEVANT PIXELS.

pose.com		chictopia.com	
α_{min}	# images	α_{min}	# images
0.05	1,308	0.05	1,257
0.09	973	0.10	937
0.13	475	0.15	421

B. Evaluation Procedure

To evaluate the accuracy of the algorithms, we have used the Jaccard distance. Specifically, given the correct set of labels $\mathcal{L}_q^*$ and the predicted set of labels $\mathcal{L}_q$ for each target image q in the test-set $\mathcal{T}$, the Jaccard distance J is given as:

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

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

We conducted five-fold cross validation in order to assess the accuracy of the evaluated algorithms. Thus, the dataset was arranged into five folds, including training, validation and test sets. At each run, three folds are used as training-set, one fold is used as validation-set (i.e., in order to build the $\gamma(\mathcal{L}_q)$ function), and the remaining fold is used as test-set. The results reported are the average of the five runs. Parameters used are those that lead to the best results for each evaluated algorithm.

C. Results and Discussion

Our experiments were devised to investigate: (i) how accuracy is impacted by exploiting textual information in addition to images, (ii) how accuracy is impacted by the proportion of relevant pixels, (iii) how accuracy is impacted by the number of base images, (iv) how M3CA algorithm performs relatively to the baseline.

Figure 6 shows accuracy numbers for the chictopia.com dataset. As can be seen, exploiting textual information in addition to image information, leads to accuracy improvements of up to 10%. As expected, accuracy increases with the number of base images, since in this case more training information is available. Slightly better accuracy numbers are obtained with higher values of α_{min}. Finally, M3CA provides accuracy improvements that vary from 2% (CLAC top-7) to 40% (CLAC top-3). The same trend is observed for pose.com, as shown in Figure 7, although the improvements upon the baseline are less impressive.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the M3CA approach which leverages the advantages of multi-instance/multi-label learning to exploit multi-modal information for clothing annotation. The label assignment is obtained by maintaining the consistency between the topics by the visual information and text information. Experiments show the advantages of M3CA over state-of-the-art algorithms. Although M3CA is designed for clothing annotation, it is possible to be applied to other multi-modal tasks. In the future, M3CA can be adapted to use dictionary learning in combination with global descriptors to improve the understanding of the image.

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

REFERENCES

[1] R. Stehling, M. Nascimento, and A. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *CIKM*, 2002, pp. 102–109.
[2] M. Unser, "Sum and difference histograms for texture classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 1, pp. 118–125, 1986.
[3] 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.
[4] J. Zegarra, N. Leite, and R. Torres, "Wavelet-based feature extraction for fingerprint image retrieval," *Journal of Computational and Applied Mathematics*, 2008.
[5] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD*, 1993, pp. 207–216.
[6] A. Veloso, W. Meira Jr., M. Gonçalves, and M. Zaki, "Multi-label lazy associative classification," in *PKDD*, 2007, pp. 605–612.
[7] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recogn.*, vol. 45, no. 1, pp. 346–362, Jan. 2012.
[8] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV*, 2002, pp. 97–112.
[9] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *NIPS*, 2003.
[10] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma, "Dual cross-media relevance model for image annotation," in *MULTIMEDIA*, 2007, pp. 605–614.
[11] Y. Yu, W. Pedrycz, and D. Miao, "Neighborhood rough sets based multi-label classification for automatic image annotation," *International Journal of Approximation Reasoning*, vol. 54, no. 9, pp. 1373–1387, Nov. 2013.
[12] A. K. Gulisong Nasierding, Grigorios Tsoumakas, "Clustering based multi-label classification for image annotation and retrieval," in *SMC*, 2009.
[13] G. Tsoumakas and I. Katakis, "Multi label classification: An overview," *International Journal of Data Warehouse and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

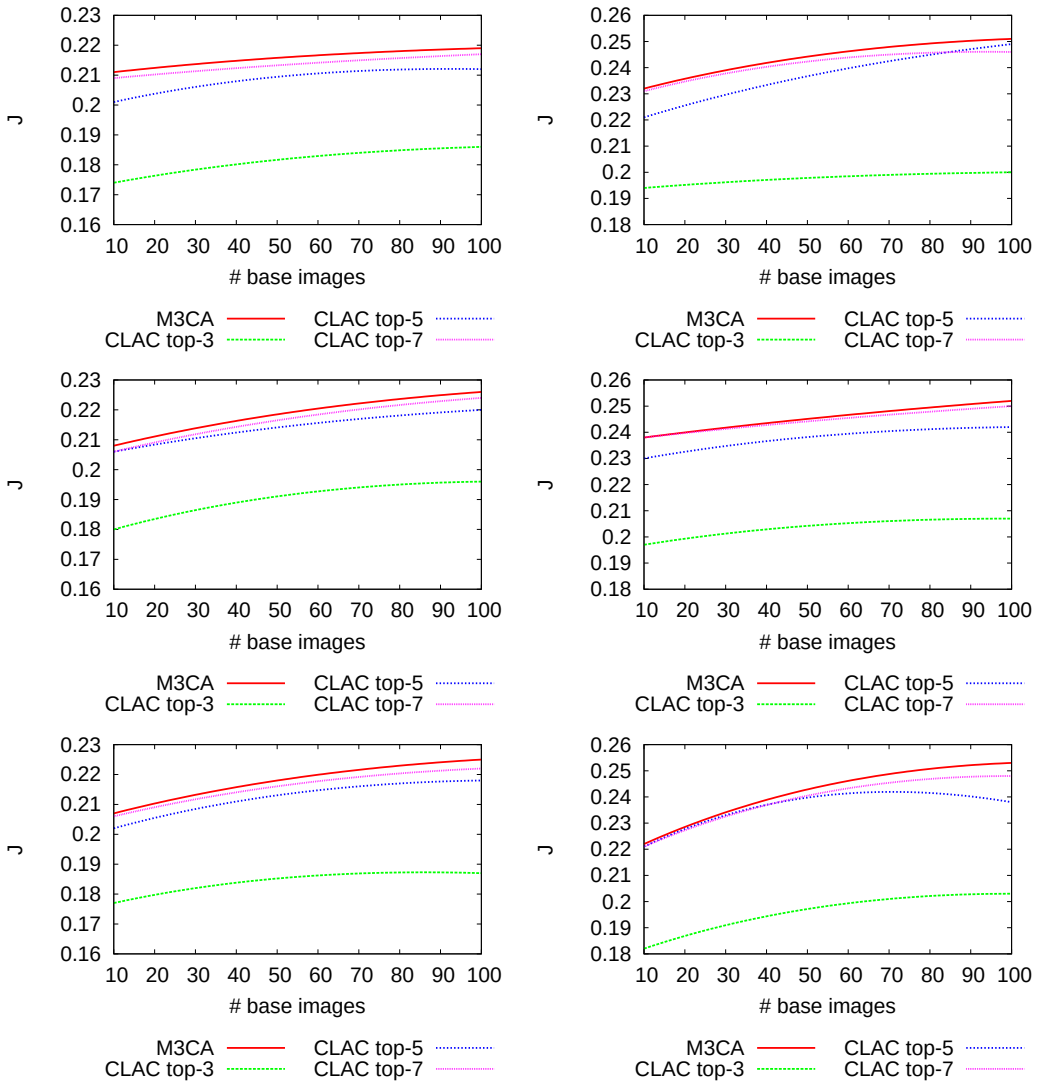

Fig. 6. chictopia.com – Figures in the left shows accuracy numbers obtained when only image information is exploited. Figures in the right shows accuracy numbers obtained when both textual and image information are exploited. First row shows accuracy numbers for $\alpha_{min} = 0.05$. Second row shows accuracy numbers for $\alpha_{min} = 0.10$. Third row shows accuracy numbers for $\alpha_{min} = 0.15$.

[14] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification." in *GrC*, 2005, pp. 718–721.

[15] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Advances in Knowledge Discovery and Data Mining*, 2004, pp. 22–30.

[16] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, Nov. 2008.

[17] J. Read, B. Pfahringer, and G. Holmes, "Multi-label classification using ensembles of pruned sets," in *ICDM*, 2008, pp. 995–1000.

[18] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," in *Machine Learning*, 2000, pp. 135–168.

[19] F. De Comité, R. Gilleron, and M. Tommasi, "Learning multi-label alternating decision trees from texts and data," in *MLDM*, 2003, pp. 35–49.

[20] A. Veloso, H. de Almeida, M. Gonçalves, and W. Meira Jr., "Learning to rank at query-time using association rules," in *SIGIR*, 2008, pp. 267–274.

[21] R. Silva, M. Gonçalves, and A. Veloso, "Rule-based active sampling for learning to rank," in *ECML/PKDD (3)*, 2011, pp. 240–255.

[22] Z. Younes, F. Abdallah, and T. Denœux, "Evidential multi-label classification approach to learning from data with imprecise labels," in *Computational Intelligence for Knowledge-Based Systems Design*, 2010, pp. 119–128.

[23] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *ICCV*, 2013, pp. 3519–3526.

[24] L.-J. Li, R. Socher, and F.-F. Li, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *CVPR*, 2009, pp. 2036–2043.

[25] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[26] B. Hasan and D. Hogg, "Segmentation using deformable spatial priors with application to clothing." in *BMVC*, 2010, pp. 1–11.

[27] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *CVPR*, 2012, pp. 3570–3577.

[28] M. Tokumaru, N. Muranaka, and S. Imanishi, "Virtual stylist project - examination of adapting clothing search system to user's subjectivity with interactive genetic algorithms," in *CEC*, vol. 2, Dec. 2003.

[29] D. Vogiatzis, D. Pierrakos, G. Paliouras, S. Jenkyn-Jones, and B. J. H. H. A. Possen, "Expert and community based style advice," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 647–10 655, 2012.

[30] E. Shen, H. Lieberman, and F. Lam, "What am i gonna wear?: Scenario-

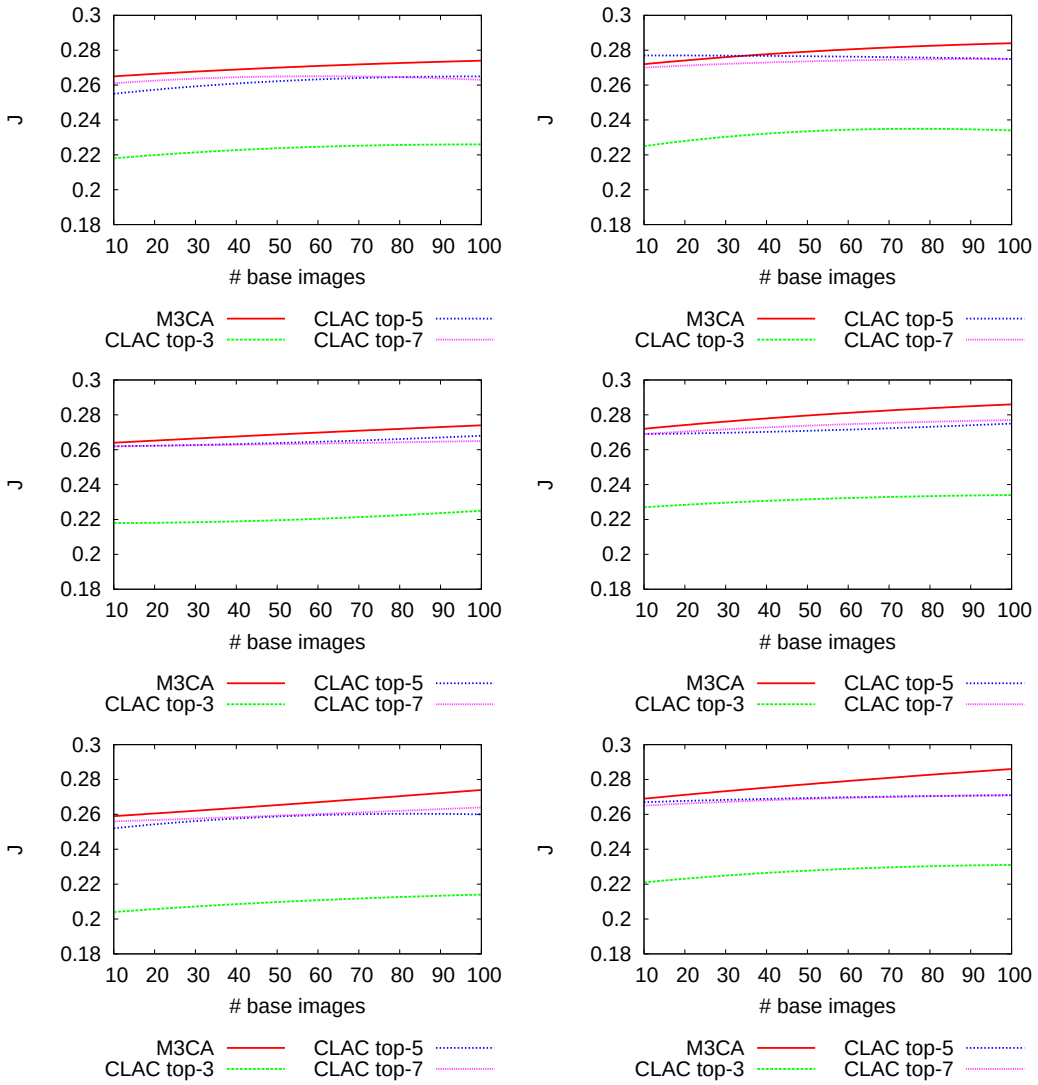

Fig. 7. pose.com – Figures in the left shows accuracy numbers obtained when only image information is exploited. Figures in the right shows accuracy numbers obtained when both textual and image information are exploited. First row shows accuracy numbers for $\alpha_{min} = 0.05$. Second row shows accuracy numbers for $\alpha_{min} = 0.09$. Third row shows accuracy numbers for $\alpha_{min} = 0.13$.

oriented recommendation," in *International Conference on Intelligent User Interfaces*, 2007, pp. 365–368.

[31] R. da Silva Torres and A. X. Falcão, "Content-based image retrieval: Theory and applications," *RITA*, vol. 13, no. 2, pp. 161–185, 2006.

[32] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, no. 1, pp. 1–19, 2004.

[33] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *CVPR*, Washington, DC, USA, 1997, pp. 762–768.

[34] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *MULTIMEDIA*, 1996, pp. 65–73.

[35] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vision*, vol. 7, no. 1, pp. 11–32, Nov. 1991.

[36] C. Huang and Q. Liu, "An orientation independent texture descriptor for image retireval," in *ICCCS*, 2007, pp. 772–776.

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

[38] R. Baeza-Yates and B. R-Neto, *Modern Information Retrieval*. Addison-Wesley-Longman, 1999.

[39] U. Fayyad and K. Irani, "Multi interval discretization of continuous-valued attributes for classification learning," in *IJCAI*, 1993, pp. 1022–1027.

[40] R. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[41] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification." in *Advances in Neural Information Processing Systems 14*, 2001, pp. 681–687.

[42] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1385–1392.