A Human vs. Machine Challenge in Fashion Color Classification

Costantino Grana, Daniele Borghesani, and Rita Cucchiara

Università degli Studi di Modena e Reggio Emilia,
Facoltà di Ingegneria "Enzo Ferrari"

Abstract. For this demo, we present a set of stark applications designed to evaluate the performance of a color similarity retrieval system against human operators performance in the same tasks. The proposed series of tests give some interesting insights about the perception of color classes and the reliability of manual annotation in the fashion context.

Keywords: Human-Machine Interaction, Image Retrieval, Color Analysis, Evaluation.

1 Introduction

Image analysis in the context of clothing and fashion can improve the user experience within Internet shopping (assisting the end-user in the process of searching the desired garments), as well as the quality control of manual annotations, impacting positively on the perception of the quality of the company itself. However properties like color or texture, as well as the very different garments available in the fashion community, are often not clearly distinguishable, even for a human operator. Color in particular is strictly related to its perception, being influenced by nuances, illumination, gamma correction of the screen etc., as well as cultural background of operators and customers, or advertising rules used to improve the product appealing neglecting objective color reproduction.

In this demo, we show simple applications for collecting color evaluations from human operators and compare them to machine outputs, exploiting also deceptive strategies to emphasize biases of operators' judgments.

2 Brief System Description

After a first background removal, the system presents two main modules. A *phototype detection* module first classifies images according to the shooting type by means of a Random Forest classification [1] using histogram projections over x and y. Then, according to the shooting type, the *Interest Garment Selection* module removes both skin and additional garments and accessories to obtain a clear picture of the object of interest: it is a segmentation problem, with an automatic initialization inferred by geometric constraints learned from the dataset. The segmentation was performed exploiting a Gaussian Mixture Models

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 631–634, 2012.

in color domain, using the iterative energy minimization approach proposed in the GrabCut algorithm [2]. This iterative procedure aims at minimizing by minimum cut the following Gibbs energy:

$$E(\alpha, k, \theta, z) = U(\alpha, k, \theta, z) + V(\alpha, z) \tag{1}$$

where α is the current segmentation mask ($\alpha_n \in \{0, 1\}$), k is a vector, with $k_n \in \{1, \dots K\}$, assigning to each pixel a unique GMM component, one component either from the background or the foreground model. θ is the set of parameters of the GMM, and z is the image pixels.

Once obtained the mask of the interesting garment, a color histogram with adaptive binning [3] (trained upon the color classes available in our dataset) is computed and used for color retrieval and classification.

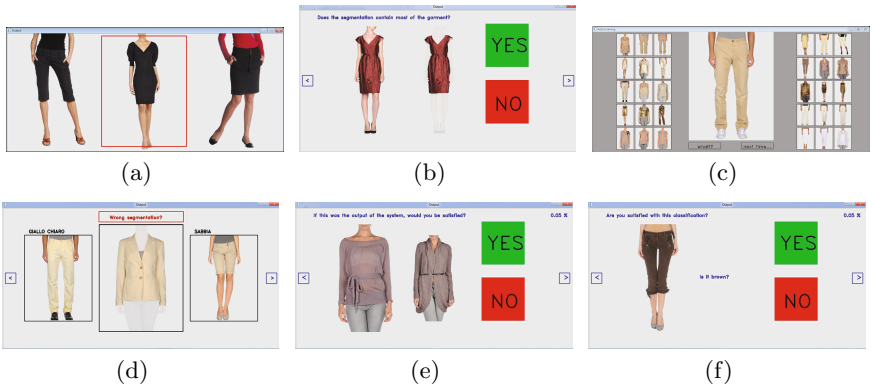
3 Discussion

In collaboration with a worldwide leader in fashion e-commerce, we collected a dataset for fashion retrieval and color classification: 60204 images of different garments and accessories, divided into mannequin shootings (23%), models (52%) and still life (25%). The fashion retailer provided a color category for each image, ranging in a set of 60 nuances unevenly distributed in the color space.

In the usual workflow our partner detailed us, the picture of the garment is taken in an arranged room; then a human operator checks the photo *and* the real garment in his hands, assigning a color category within the company database. If we analyze this entire situation, we can easily highlight a number of inaccuracies. The initial photo is taken in a controlled but not calibrated context, so we cannot access to the sRGB color values: a color checker would solve the problem, but this requires a change in the production workflow that, even if endorsed by the company, would not solve the problem with the thousands of images already taken. An operator assigns a color category in a different working place, with a different environmental illumination and therefore a different perception of the color, either for the real thing and the digital reproduction on his computer screen. Moreover, given the enormous amount of garments processed every day, usually it is a one-operator-only opinion.

At last, we end up dealing with a highly confused annotation, and the problem is particularly relevant for two reasons. Firstly the efficiency of color description: we need to know even a very small but as-objective-as-possible subset to train the color classification with; secondly, performance evaluation: we cannot evaluate effectively its performance if a significant part of the dataset is inconsistent.

The problem became initially clear collecting data using the application in Fig. 1(a). Given an initial query depicted in the center of the view, we asked to select which one is believed the most similar. On its left and on its right, randomly, the nearest neighbor within the original color class and the one within the entire dataset is depicted. In the 55% of cases the operator chooses the automatic classification, while in the remaining 45% he preferred the original annotation, with a consistent drop in favor of automatic annotation if the median



(a) (b) (c)

(d) (e) (f)

Fig. 1. Screenshots of the very minimal applications proposed in this demo. The rude design is due to our prevailing testing purposes as well as the need to adapt to our partners' production context.

of the query class is presented (78% against 22%). The interclass variance is therefore considerable.

The annotation correction process started asking operators a very limited set $\tilde{R}$ of samples for each nuance. This procedure was completely manual, and led to the definition of a pantone: the idea is to substitute the usually adopted set of uniform color patches with a set of full garments, thus holding a more comprehensive description of color. Then we selected an initial subset of 10000 images (*color reference*, R) whose segmentation mask correctness has been manually verified (using UI of Fig. 1(b)). We employed an active learning framework [4] to assist the user in this task: for each combination of features and classifiers we tested (GMM, dominant color and RGB histogram with Random Forest or SVM, etc...), we performed a first rank of the results based on the feature distance, then a re-rank based on the difference between the first-ranked and the second-ranked elements. Using the interface in Fig. 1(c), with the query at the center and samples from $\tilde{R}$ around, we therefore asked operators' opinion for those garments in which the classifier is more confused about. By keeping only results with higher level of concordance between operators, we obtained a cleaner dataset of 5029 images.

For the final refinement, using again the interface of Fig. 1(a), we asked opinions to operators, but this time we presented them only samples coming from the R, logging how many times a single sample is used in a choice concordant to the automatic segmentation Q_A and instead how many times it is used in a choice concordant to the original annotation Q_Y. We found out that only a fraction of the color reference is actually used (44%, 2258 images), and for each image I a reliability factor C_I has been computed as:

$$C_I = Q_A / (Q_A + Q_Y) \tag{2}$$

Setting an empirical threshold of 0.7, we could reduce the color reference R' down to 897 images.

R' provided a fairly objective reference to train the adaptive histogram descriptor and to be used to assign and evaluate color annotations. Using the interface of Fig. 1(d) we can therefore compare the human annotation against the machine one: given a query in the center, the operator selects the most similar color among the alternatives; in addition, if the user chooses a sample coming from the original annotation, the interface asks if the competitor would be anyway acceptable. 13.3% strongly agrees with the original annotation, 45.7% strongly agrees with the automatic one and the remaining 50%, despite agreeing with the original, considers as acceptable the automatic one.

Finally, we tried to deceive the operators with the interfaces in Fig. 1(e) and Fig. 1(f). We told them that the application was simply a mean to collect a "second opinion" and correct potential errors: it was a half lie, as only half of the displayed results actually belonged to the original annotation, while the others were the outputs of the automatic system. The operators were thus implicitly biased towards the displayed results believing them all the original annotation. Averaging the results on different random subsets and operators, we found out that they agree with themselves in 74.5% of the cases, while they agree with the automatic system similarly in 71.1% of the cases.

4 Conclusions

In our experience, the automatic classification by color in the fashion context, even if based on the digital representation only and without color calibration, is capable of reaching the same reliability of a human operator employed in the same task. Despite the problem remains strongly ill-posed, or perhaps precisely because of this, we can conclude that a good segmentation algorithm and a good color feature can compete with humans in the process of color annotation in such datasets, guaranteeing at the same time a dramatic reduction of the processing times within the company workflow.

References

1. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
2. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": interactive foreground extraction using iterated graph cuts. In: ACM SIGGRAPH, pp. 309–314 (2004)
3. Grana, C., Borghesani, D., Cucchiara, R.: Class-based color bag of words for fashion retrieval. In: IEEE International Conference on Multimedia and Expo., Melbourne, Austrialia (2012)
4. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–12 (1994)