Learning Facial Attributes by Crowdsourcing in Social Media

Yan-Ying Chen
National Taiwan University
yanying@gmail.com

Winston H. Hsu
National Taiwan University
winston@csie.ntu.edu.tw

Hong-Yuan Mark Liao
Institute of Information Science, Academia Sinica
liao@iis.sinica.edu.tw

ABSTRACT

Facial attributes such as gender, race, age, hair style, etc., carry rich information for locating designated persons and profiling the communities from image/video collections (e.g., surveillance videos or photo albums). For plentiful facial attributes in photos and videos, collecting costly manual annotations for training detectors is time-consuming. We propose an automatic facial attribute detection method by exploiting the great amount of weakly labelled photos in social media. Our work can (1) automatically extract training images from the semantic-consistent user groups and (2) filter out noisy training photos by multiple mid-level features (by voting). Moreover, we introduce a method to harvest less-biased negative data for preventing uneven distribution of certain attributes. The experiments show that our approach can automatically acquire training photos for facial attributes and is on par with that by manual annotations.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision - Applications

General Terms

Algorithms, Experimentation

Keywords

crowdsourcing, facial attribute, social media

1. INTRODUCTION

Besides low-level features for face recognition, the rich set of facial attributes (e.g., gender, race, age, hair style, smile, etc.) has been shown promising for describing target persons or profiling human activities [2]. For example, we can locate a designated person in surveillance videos by (composite) facial attributes (e.g., a bearded man). For market research, we can understand the user's preferences or profiles by analyzing the facial attributes in her photo albums; for example, "most of Alice's friends are Asian girls with long hairs." Prior works for automatic facial attribute detection solely relied on supervised learning with manually collected training photos from limited sources. Moghaddam et al. [3] proposed a gender classification approach using FERET, a standard face benchmark. Laboratory data generally fails to cover the rich facial appearances in consumer photos. Kumar et al. [2] trained 10 attributes by manually labelling certain images collected from the internet. *Crowdsourcing* – exploiting enormous user-contributed images and tags from the web – has been shown promising for easing the pains for collecting the vital annotations for supervised learning. For example, Fergus [1] preliminarily introduced internet images for image classification. Ni et al. [4] constructed an age estimator using web images and text information.

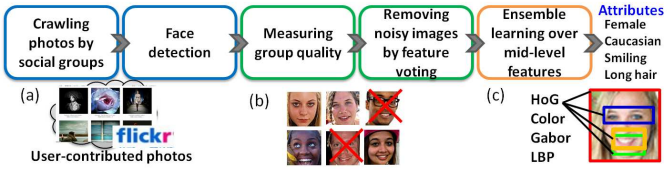

Figure 1: Leveraging photos from social groups for training facial attributes. (a) Photos from the groups. (b) Rejecting noisy images by group quality and (mid-level) feature voting. (c) Ensemble learning by the automatically collected training images.

Although crowdsourcing tends to be effortless, the quality for annotations (images) is questionable. Different from the previous works for crowdsourcing, we leverage the social groups from the community-contributed photo services (e.g., Flickr). It is natural since such photos are rich in facial attributes across races and countries. Meanwhile, we locate semantic-consistent photos from the designated user groups retrieved by keywords for the target facial attributes. To ensure annotation quality, we propose an quality measurement to rank social groups of proper semantic correlation and a voting scheme (by multiple mid-level features) to reject noisy training photos. Furthermore, we present an approach to harvest less-biased negative training images to prevent uneven distribution in certain attributes. We show in Section 3 that our approach successfully rejects noisy images for training facial attributes and the accuracy is competitive with learning by manual annotations.

2. LEARNING FACIAL ATTRIBUTES FROM SOCIAL GROUPS

Learning facial attributes confronts two difficulties: (1) current training data cannot cover the rich diversities of facial attributes ,(2) learning numerous attributes requires huge annotation efforts. To address the problems, we crowd-

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

(a) (b)

	Group name	# of photos	# of faces	faces/photos (%)
	Big Woman	1470	387	26.33
*	Cat women	2970	33	1.11
	SOUL WOMAN OF THE MONTH	870	237	27.24
*	Women in the Arts	4470	182	4.07
	Women of india	4470	1017	22.75
	The Woman Adored	3870	776	20.05

(c)

Figure 2: Query results in Flickr: (a)Image-level query by "man" (b)Group-level query by "man" (c) Two low-quality social groups (for cats and arts) for facial attribute woman.

source photos from community-sharing services since they are abundant, effortless, diverse, and up-to-date. For the quality issue, we crawled the related photos from the social groups with high quality measures and then reject the noisy training photos by mid-level feature voting (cf. Fig. 1).

Crawling photos from social groups – Social groups accommodate user-contributed photos of the same theme. As shown in Fig. 2(a)(b), group-level query introduces much less noisy photos than image-level query. We retrieve a group list by the keywords pertaining to the facial attributes (e.g., woman). Faces from these photo groups are detected.

Measuring group quality – Finding relevant groups by keywords suffers from text ambiguity. For example, the group "Cat women" shares cat photos, which are less relevant to the facial attribute, "women." We observe that such irrelevant groups containing fewer faces (cf. Fig. 2(c)). We then take the ratio of face numbers among the whole photos in the group as a quality measure for each group. We select photos from the high quality groups as the candidate (positive) training photos and also balance the number of faces from different groups for ensuring diversity.

Removing noisy training photos by feature voting – Noisy photos still exist because users tag a photo rather than a face. For instance, the tag "woman" or "man" both incurs noisy photos if they contain a woman and a man. We propose a voting scheme by multiple mid-level features to filter out such noisy photos (cf. Fig. 1(c)). Each mid-level feature is a SVM classifier with varying low-level features (i.e., Gabor filter, Histogram of Gradient, Color, Local Binary Patterns) extracted from each face component (i.e., whole face, eyes, philtrum, mouth). Such mid-level features are rich and carry semantic meanings for facial attributes as shown in [2]. We then reject the candidate training photos with low voting scores. Note that such mid-level classifiers are trained by 5-fold cross-validation from the training photos.

Harvesting less-biased negative data – For facial attribute classification, randomly picking examples as negative data is impractical because attributes are usually correlated (e.g., "man" and "beard"). We propose two solutions for collecting negative training photos. We use antonyms of the attribute to exploit negative data if the antonyms are explicit. Otherwise, we seek *universal background data* (UBD), a background training data collected by neutral words such as "people," "persons" that are not specific to any attributes. For each attribute, we also reject photos in UBD that are strongly correlated to (by voting) the mid-level features since they might be related to certain attributes. UBD reflects the background photos of facial attributes in consumer photos so that it is less biased than defining specific antonyms for negative training photos. We observed that UBD is effective for the attributes not frequently observed in photos (e.g., beard, sunglasses, etc.). The assumption is reasonable since the attributes appeared frequently (e.g., gender, age) often have explicit antonyms for acquiring negative data.

Ensemble learning over mid-level features – Provided the crowdsourced training photos, we then adopt Adaboost to aggregate the rich set of mid-level features to construct each facial attribute detector [2].

Table 1: Detection accuracy (%) for four attributes. Our approach (c) is competitive with that by manual annotations (a) and outperforms that by noisy photos specified by (image-level) keywords only (b).

method	gender	smile	kid	beard	beard(UBD)
(a) manual	85.63	87.04	87.88	85.17	-
(b) image-level	78.08	80.24	74.21	73.67	-
(c) our approach	83.02	86.28	85.97	81.70	82.97

3. EXPERIMENTS AND DISCUSSIONS

Tackling noisy training photos – We compare our approach with (a) learning by manual annotation [2] and (b) learning by noisy images from the web [1]. Each attribute is trained by 1800 positive faces and 1800 negative ones and evaluated on 400 faces. The face images are harvested from Flickr. Table 1 shows the impact of the proposed method. The accuracy by automatically crowdsourcing the social groups and rejecting noisy photos reaches almost similar performances across facial attributes learned by manual annotations. Intuitively, simply crawling by image-level keywords degrades the quality of training photos and thus yields lower accuracy.

Effectiveness of UBD – For preliminary understanding, we experiment on "beard" attribute since it is not frequently observed in photos and lacks explicit (visual) antonyms. Although antonyms likely discriminate the positive from negative facial attributes, it possibly causes data skew if the antonyms cannot cover the whole attribute space. For example, data is unevenly distributed if we only define "woman" and "kid" as the antonyms for "beard." As Tabel 1(c) shows, UDB (82.97%) outperforms that by negative training photos acquired by antonyms (81.70%) since UBD avoids data skew and does help classification.

Leveraging photos in social groups and removing noisy photos by mid-level feature voting, we can automatically acquire effective training photos for facial attribute detection. We also propose UBD to obtain less biased negative training photos. We aim to extend the crowdsourcing methods to a huge set of facial attributes and enable applications (e.g., user profiling and retrieval) by the rich facial attributes.

4. REFERENCES

[1] R. Fergus et al. Learning object categories from google's image search. *In IEEE ICCV*, 2005.

[2] N. Kumar et al. Facetracer: A search engine for large collections of images with faces. *In ECCV*, 2008.

[3] B. Moghaddam et al. Learning gender with support faces. *In IEEE TPAMI*, 2002.

[4] B. Ni et al. Web image mining towards universal age estimator. *In ACM MM*, 2009.