Cuteness Recognition and Localization in the Photos of Animals

Yu Bao
School of Software
Dalian University of Technology
baoyu_dlut@foxmail.com

Jing Yang
College of Arts
Chongqing University
yangjing19900530@163.com

Liangliang Cao
IBM Watson Research Center
liangliang.cao@us.ibm.com

Haojie Li
School of Software
Dalian University of Technology
hjli@dlut.edu.cn

Jinhui Tang
School of Computer Science
Nanjing University of Science and Technology
tangjh1981@acm.org

ABSTRACT

Among the flourishing amount of photos in the social media websites, "cute" images of animals are particularly attractive to the Internet users. This paper considers building an automatic model which can distinguish cute images from non-cute ones. To make the recognition results more interpretable, a lot of efforts are made to find which part of the animal appears attractive to the human users. To validate the success of our proposed method, we collect three new datasets of different animals, i.e., cats, dogs, and rabbits with both cute and non-cute images. Our model obtains promising performance in distinguishing cute images from non-cute ones. Moreover, it outperforms the classical models with not only better recognition accuracy, but also more intuitive localization of the cuteness in the images. The contribution of this paper is three-fold: (1) We collect new datasets for cuteness recognition, (2) We extend the powerful Fisher Vector representation to localize cute part in the animal recognition, and (3) Extensive experimental results show that our proposed method can recognize cute animals of cats, dogs, and rabbits.

Categories and Subject Descriptors

I.5.4 [**Applications**]: Computer Vision; H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

cute images; recognition; localization; cats; dogs; rabbits

1. INTRODUCTION

Dogs, cats, and rabbits are close friends of human being and their photos are very popular in social media. They are among the most popular subjects in social media websites. People are especially interested in "cute" photos of these animals. For example, *Maru*, a cat from Japan has resulted in more than 165 million views on its Youtube channel [1]. Cute photos stimulate positive feeling and intensive emotion and there are many studies in psychology on the "cuteness" of images. Some recent research work [11, 15] showed the power of cuteness and suggested that viewing cute images can promote focused attention and happy emotions.

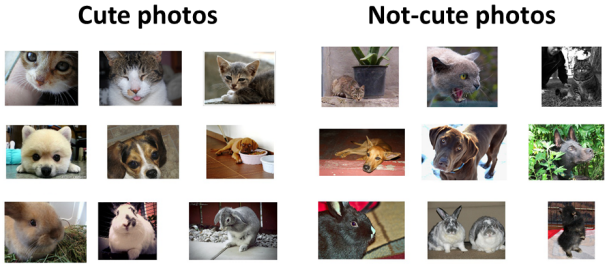

Figure 1: Some examples of our new dataset of cuteness recognition.

In this paper, we attempt to take a computational approach to understand the cuteness of animal photos. To this end, we downloaded photos from Flickr[2] and asked users to judge whether a photo is cute or not. We selected images according to users' consensus and obtained a dataset with 3,600 photos of dogs, cats and rabbits. The ratio of cute and non-cute photos in this dataset is about 1 : 2. Figure 1 shows some example photos from our dataset.

To learn a statistical model of cuteness from this dataset, we looked into extensive literature on general object recognition. In prevailing object recognition systems, there are two steps to train a classifier: (1) Image description, which extracts features of image and then encode the features to image level representation. (2) Statistical learning, which employs machine learning approaches to train a classifier based on image representation. In the popularly-used effective Bag-of-Visual-Words (BoVW) model [1], the image de-

MM'14, November 03 - 07 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2647868.2655046.

[1]http://techland.time.com/2012/06/27/a-brief-historyof-cats-on-youtube/
[2]https://www.flickr.com/

scription is the histogram of Scale Invariant Feature Transform (SIFT) [10] features, and the statistical model usually uses Support Vector Machine (SVM) to train nonlinear classifiers from high dimensional patterns.

Bag-of-visual-words model has many advantages in solving general object recognition. However, it cannot satisfy the requirement of our cuteness detection problem. We prefer a new model with the following capabilities:

- The model is able to take spatial context into account, and use the neighborhood information to infer whether the subject is cute or not.
- The model could not only determine whether an image is cute or not, but also find which part of the animal is important for inferring .

Traditional bag-of-words model could not satisfy these requirements since it overlooks the spatial information from images. Thus, a reliable and efficient solution needs be developed instead.

Our proposed solution is based on Fisher vector (FV) [14], which employs Gaussian Mixture Model (GMM) to build a parametric generative model for local features. However, classical FV model could only give image-level estimation but is not able to find the cute parts. We extend the previous work on FV and obtain promising results on our dataset. We will explain our method in this paper and report the performance of our method.

To the best of our knowledge, there is no previous work on the cuteness recognition of images. However, there are still a number of previous studies on animal images which are helpful to this study. Many research studies have been carried out on object recognition and detection from cluttered background [2, 3, 7, 12, 20]. A recent research focus is to model highly deformable of animals [8, 13, 17], and another focus is to discriminate different breeds of animals and to identify fine-grained categories [4].

This paper is organized as follows. Section 2 explains the proposed cuteness model in details. Section 3 evaluates the performance of the proposed method on cuteness recognition and localization. We conclude this paper in Section 4.

2. THE PROPOSED METHOD

The cuteness of animal images involves the subtle description of the image. Following previous work, we employ dense SIFT [10] to represent the local features of the animal. We dense sample patches with a stride of step pixel and compute the SIFT features for each patch. The process is repeated at five scales, with a scaling factor. Each SIFT feature is a 128 dimensional descriptor. The next step is how to encode dense features into image level representation.

Traditional bag-of-words model suggests to convert a SIFT vector to an index number which corresponds to an existing visual word in the vocabulary. Though lots of successes have been demonstrated in image retrieval [9], image recognition [16], etc, this method generates a sparse representation and thus lots of information of local descriptor is lost. To keep more information, we follow the Fisher Vector [14] for image representation to reduce the loss of information in BoVW.

Fisher Vector is a method that summarizes a varying number of local feature descriptors (e.g. SIFT). It first fits the Gaussian Mixture Model (GMM) which is a parametric generative model to the set of local features, and then represents the image with the derivatives of the log-likelihood of the model by encoding them with respect to the model parameters [5]. Following [14] ,we train a GMM using the extracted SIFT features and get the parameters Θ = (μ_k,σ_k,π_k:k=1,2,...,K). μ_k,σ_k,π_k is mean, covariance and the mixture weight of the GMM. K is the number of Gaussians in the GMM. Let I=(x_1,...,x_N) be a set of d dimensional SIFT features extracted from an image. The GMM associates each vector x_i to a mode k in the mixture with a strength given by the posterior probability:

$$q_{ik} = \frac{exp[-\frac{1}{2}(x_i - \mu_k)^{\mathrm{T}}\sigma_k^{-1}(x_i - \mu_k)]}{\sum_{t=1}^{K} exp[-\frac{1}{2}](x_i - \mu_t)^{\mathrm{T}}\sigma_t^{-1}(x_i - \mu_t)]}$$

For each mode k, consider the mean and covariance deviation vectors:

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{K} q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}}$$

$$v_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{K} q_{ik} [(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}})^2 - 1]$$

where $j = 1, 2, ..., d$ spans the vector dimensions. The FV of image I is the stacking of the vectors u_k and then of the vectors v_k for each of the K modes in the Gaussian mixtures:

$$\Phi(I) = [...u_k...v_k...]$$

Thus, the FV dimensionality is $2Kd$, where K is the number of Gaussians in the GMM, and d is the dimensionality of the feature vector.

Previous studies employ linear SVM to model Fisher Vectors. Suppose w is the weight vector and b is the bias of SVM, the score of an input x is

$$s = w'x + b$$

Here x represents image feature and s denotes whether this image is cute or not.

We make two moderate extensions on the base of Fisher vectors:

1. We use the spatial information of local descriptor to find the cute part and model a cute part as the neighborhood of ellipse shape.
2. Instead of using SVM directly, we use an ensemble method to select the most useful components for cute animals.

For the first extension, we follow the work in [14] by appending the feature coordinates to the visual features. Thus our dense features have a form like [$SIFT_{xy}$,$\frac{x}{W}$-$\frac{1}{2}$,$\frac{y}{H}$-$\frac{1}{2}$], where $SIFT_{xy}$ is the SIFT descriptor at (x,y), and W and H are the width and height of the image. After training a GMM we get the parameters Θ = (μ_k,σ_k,π_k:k=1,2,...,K). μ_k and σ_k are the mean and diagonal covariance for k-th Gaussian Model. By looking at the last two dimensions from each Gaussian component, we are able to model the spatial information of the subject. We can interpret the mean vector as the center of an ellipse and use the diagonal covariance to show the skewness of the ellipse. Note that the position of the ellipse is different from image to images. The Fisher vector representation will pool local patches for every image and compute the location adaptively.

Figure 2: Localize cuteness using the ellipse detection. Left: local patches ranked by classical FV model. Right: ellipses detected by our ensemble model. The blue points indicate SIFT features belonging to the top local patches with high weights. The ellipses correspond to selected "weak" classifiers in Adaboost.

A challenging task along with image-level cuteness recognition is to localize which part of the animal is cute, *i.e.*, to find out which component of GMM corresponds to the cute animal. We makes it possible to evaluate how important certain Gaussians is by using the Adaboost algorithm. The Adaboost [6] is an algorithm for constructing a "strong" classifier as linear combination of "weak" classifiers. Suppose the FV dimensionality is $2Kd$, where K is the number of mixtures in the GMM, and d is the dimensionality of local patch features plus two dimensional coordinates. The weak classifier is a linear SVM model trained on $2d$ features for every component in the FV.

Let $h_t(x)$ be the weak classifier, D_t be the weight of train image and y be the label of train image. Then we start a loop for computing the important Gaussians. First, the weight of every train image is same and we consider selecting a weak classifier with the smallest weighted error,

$$h_t = arg \min_{h_j \in H} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$$

where m is the number of training images and H is the set of weak classifiers.

Second, we compute α to update D_t

$$\alpha_t = \frac{1}{2} log(\frac{1+r_t}{1-r_t}),$$

where

$$r_t = \sum_{i=1}^{m} D_t(i) h_t(x_i) y_i$$

Note that α_t is the weight of the weak classifier which is proportional to r_t, the accuracy of the classifier. Then we update D_t using α_t.

$$D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t},$$

here Z_t is a normalization factor to make D_{t+1} is a distribution. To select more components, we iterate the above procedure and find more h_t. The final strong classifier is

$$H_x = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

By combining multiple "weak" classifiers, we can select the GMM components corresponding to the cute animals. Figure 2 shows an example of cuteness localization. It is easy to see the difference between our method and the traditional FV. Traditional FV yields a unified SVM using all the components with the max margin criteria. The resulted SVM optimizes the image level estimation, but fires on too many local patches. In contrast, our ensemble method only picks the top components as the localization of the cute parts.

3. EXPERIMENTS

Table 1: Accuracy of recognizing cute images using our method and traditional bag-of-words model.

	cat	dog	rabbit
BOW+SVM	87.23%	75.81 %	75.95 %
Our method	89.36%	79.78 %	77.04 %

Novel Cuteness Datasets. We collected new datasets with three categories: cat, dog, and rabbit. Each image is labeled with either "cute" or "non-cute". The cat category contains 1826 images with 1298 non-cute and 528 cute images. The dog category contains 1477 images with 987 non-cute image and 490 cute images. The rabbit category contains 1143 images with 761 non-cute image and 372 cute images. For the cat dataset we randomly selected 1100 non-cute images and 428 cute images as training set and the remaining images were used for testing. In the dog dataset, 800 non-cute images and 400 cute images were selected as training set and the remaining images were used as testing set. In rabbit dataset we selected 650 non-cute images and 500 cute images as training set and the remaining images were used as test set.

Experimental Settings and Results. The experiments were carried out by testing three methods on the three categories. We used the following settings: the number of visual word is 512; the number of Gaussian components is 512. We also applied the PCA to SIFT, reducing its dimensionality from 128 to 80. After encoding spatial information, the feature dimension becomes $d = 82$. We trained three cuteness classifiers for the three categories respectively.

To detect the cute parts of interests, we trained an ensemble model with 20 weak classifiers. Each classifier corresponds with a component in the Fisher vector. According to the last two dimensions in each component, we can identify a ellipse region in the image. We plotted these regions if the weak classifier gives positive score. Figure 3 illustrates the detection results. It is easy to see that our model could reliably locate the cute parts of the animal images.

We compared the performance of our cuteness recognition method with the classical bag-of-words (SIFT + SVM) model in Table 1. We can see that our method consistently outperforms the BOW model in all the three datasets. Moreover, the satisfying accuracies (77%-89%) suggest that we can obtain a reliable model for cuteness recognition.

Table 2: Comparing our method and classical Fisher vector.

	mean Acc	number of components
Our method	82.06%	20
Fisher Vector [14]	84.31%	512

Since our model can be viewed as an extension of Fisher vector [14], it is worth comparing the performance of our

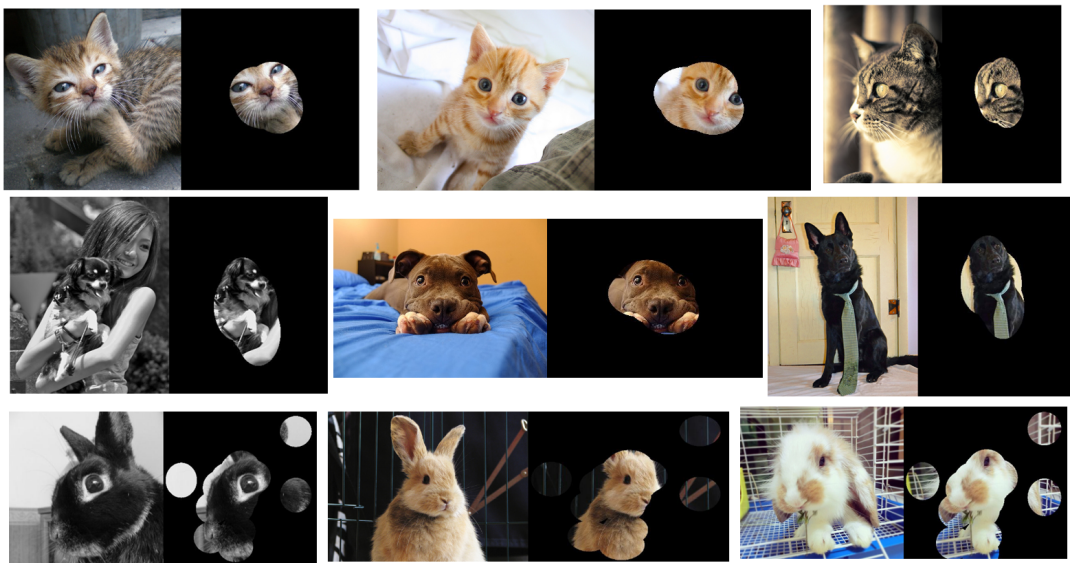

Figure 3: Cuteness localization by our method.

method and FV. As Table 2 shows, our method uses only 20 components, while FV uses 512 components. However, we obtain a comparable accuracy even with a much smaller number of components. The benefits of using less components are validated by Figures 2 and 3, where our model can localize cute parts which can be interpreted by human.

4. CONCLUSIONS

This paper proposes to recognize cute one from non-cute animals, including cats, dogs, and rabbits. We extend the classical Fisher vector method so that our new method can not only determine whether an image is cute or not, but also detect the cute parts in the images. Compared with previous work, the proposed model works reliably and generates clear localization results. In the future, we will apply our method to large scale cuteness detection. Another research interest is to investigate the attribute approach [19, 18] in cuteness recognition.

5. ACKNOWLEDGEMENTS

This work was partially supported by the Natural Science Foundation of China (61173104), National Basic Research Program of China (2014CB347600), and Open Projects Program of National Laboratory of Pattern Recognition.

6. REFERENCES

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, 2004.

[2] F. Fleuret and D. Geman. Stationary features and cat detection. *Journal of Machine Learning Research*, 9(11), 2008.

[3] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

[4] ImageNet. Fine-grained classification of dogs. http://www.image-net.org/challenges/LSVRC/2012/, 2012.

[5] T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.

[6] J. M. Jan Ŝochman. Adaboost. *Centre for Machine Perception, Czech Technical University, Prague.*

[7] L. Ladickỳ, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010.

[8] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR* , 2009.

[9] H. Li, X. Wang, J. Tang, and C. Zhao. Combining global and local matching of multiple features for precise item image retrieval. *Multimedia systems*, 19(1):37–49, 2013.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] H. Nittono, M. Fukushima, A. Yano, and H. Moriya. The power of kawaii: Viewing cute images promotes a careful behavior and narrows attentional focus. *PloS one*, 7(9):e46362, 2012.

[12] O. M. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman. The truth about cats and dogs. In *ICCV*, 2011.

[13] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *CVPR*, 2012.

[14] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[15] R. Sprengelmeyer, D. Perrett, E. Fagan, et al. The cutest little baby face a hormonal link to sensitivity to cuteness in infant faces. *Psychological Science*, 20(2):149–154, 2009.

[16] S. Tang, Y.-T. Zheng, Y. Wang, and T.-S. Chua. Sparse ensemble learning for concept detection. *Multimedia, IEEE Transactions on*, 14(1):43–54, 2012.

[17] P. Welinder, S. Branson, T. Mita, Cet al. Caltech-ucsd birds 200, 2010

[18] F. X. Yu, L. Cao, R. S. Feris, et al. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[19] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *ACM Multimedia*, 2013.

[20] W. Zhang, J. Sun, and X. Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, 2008.