Learning Multi-view Deep Features for Small Object Retrieval in Surveillance Scenarios

Haiyun Guo[1], Jinqiao Wang[1], Min Xu[2], Zheng-Jun Zha[3], and Hanqing Lu[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100190
[2]GBDTC, School of Computing and Communications, University of Technology, Sydney, Australia
[3]Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China
{haiyun.guo, jqwang, luhq}@nlpr.ia.ac.cn, min.xu@uts.edu.au, junzzustc@gmail.com

ABSTRACT

With the explosive growth of surveillance videos, object retrieval has become a significant task for security monitoring. However, visual objects in surveillance videos are usually of small size with complex light conditions, view changes and partial occlusions, which increases the difficulty level of efficiently retrieving objects of interest in a large-scale dataset. Although deep features have achieved promising results on object classification and retrieval and have been verified to contain rich semantic structure property, they lack of adequate color information, which is as crucial as structure information for effective object representation. In this paper, we propose to leverage discriminative Convolutional Neural Network (CNN) to learn deep structure and color feature to form an efficient multi-view object representation. Specifically, we utilize CNN trained on ImageNet to abstract rich semantic structure information. Meanwhile, we propose a CNN model supervised by 11 color names to extract deep color features. Compared with traditional color descriptors, deep color features can capture the common color property across different illumination conditions. Then, the complementary multi-view deep features are encoded into short binary codes by Locality-Sensitive Hash (LSH) and fused to retrieve objects. Retrieval experiments are performed on a dataset of 100k objects extracted from multi-camera surveillance videos. Comparison results with several popular visual descriptors show the effectiveness of the proposed approach.

Categories and Subject Descriptors

I.4.10 [**Image Representation**]: Multidimensional; I.5.4 [**Applications**]: Signal processing

Keywords

Object retrieval; Object representation

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806349.

1. INTRODUCTION

Nowadays an increasing number of surveillance cameras are installed in public places and produce massive video data every day. Therefore, efficiently retrieving objects of interest across large scale surveillance videos has become a hot spot of research. Although it has been studied for many years, small object retrieval in surveillance videos remains a challenging topic, primarily due to the following reasons. Firstly, the objects extracted from surveillance videos are usually of small size, even imprecise. Secondly, the image including object of interest also inevitably involves some background pixels. Thirdly, there are various complex conditions existing in surveillance areas, such as illumination changes, shadows, viewpoint shifts and partial occlusions.

Therefore, an efficient object representation is crucial to object retrieval in surveillance videos. Over the past few decades, a series of hand-crafted features have been proposed to bridge the semantic gap between image pixels and semantic concepts. One is global descriptor, which describes the coarse global color, texture or shape information distribution of input image. Yang and Yu [19] made use of color histograms and three different texture descriptors to real-time search eight kinds of clothes in surveillance videos. Calderara *et al.* [3] presented a global descriptor by estimating the color probability distribution with a mixture of Gaussians to conduct person retrieval in multi-camera surveillance scenarios. However, these global descriptors fail to distinguish between the object and unrelated pixels within image, thus introducing noise to object representation. Besides, they are not robust to some complex variations. Another is local descriptor such as SIFT [10] and GLOH [12], all of which are based on feature point detection. However, as mentioned before, the objects detected from surveillance videos are usually rather small, so there are often few even no feature points detected from an object. As a result, local descriptors extracted from small objects usually lack of adequate discrimination. Apart from the above descriptors, other researchers utilized various attributes to describe visual objects. Thornton *et al.*[16] proposed a generative model covering some attributes such as gender, hair/hat color, clothing color and bag position for person search. Although attribute features convey middle-level visual information, the extraction and fusion of various attributes are a very challenging task for small objects in surveillance videos.

With the rapid development of deep learning, it has been verified in [8, 20] that features learned by deep models are more efficient than traditional hand-crafted features. Hinton

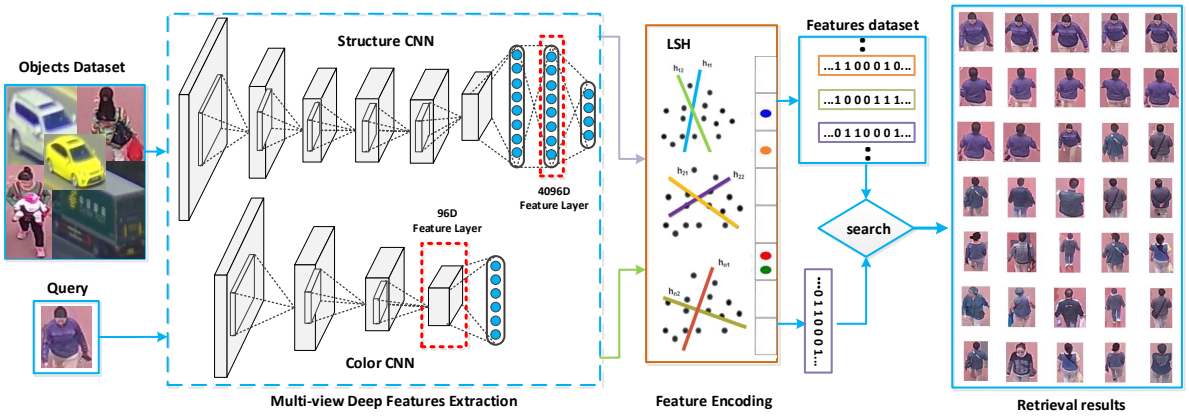

Figure 1: The overall framework of learning multi-view deep features for small object retrieval.

et al.[7] and Teng *et al.*[15] both applied features learned by deep auto-encoders to content-based similar image retrieval and obtained better retrieval performance than using pixel-based features. But as pointed out in [8], the retrieval results can have similar edge patterns but may not be semantically related. In contrast, deep representations learned by Convolutional Neural Network (CNN) conveyed rich semantic information and obtained state-of-the-art image retrieval performance in [13].

Structure and color attributes are two most crucial properties for effective object representation. In this paper, we leverage multi-view information abstracted with CNN models to jointly describe the structure and color properties of visual objects. The overall framework of the proposed approach is illustrated in Figure 1. Specifically, we take advantage of the CNN model trained on ImageNet [20] to extract deep feature from the view of structure. This kind of feature has two advantages: for one thing, it includes rich local structure information and is robust to pose change, object deformation as well as partial occlusion; for another, it delivers rich semantic information about object category. However, the above deep structure feature turns out to lack of adequate color information in retrieval experiments. Thus, we build a CNN model supervised by 11 color names [18] to abstract deep feature from the view of color, which could capture the common color property across different illumination conditions. Therefore, with discriminative CNN models, we can extract efficient and robust multi-view deep features for small objects. Furthermore, to accelerate the retrieval efficiency, the above two deep features are encoded into 256-bit binary codes by Locality Sensitive Hashing (LSH) [4] separately. Finally, we combine the complementary multi-view deep features using a late fusion strategy and obtain state-of-the-art performance for small object retrieval in surveillance scenarios.

2. MULTI-VIEW DEEP REPRESENTATION

Since introduced by LeCun [9] in the early 1990's, CNN has demonstrated record-beating performance at challenging tasks such as image classification, object detection and face recognition. Three crucial architectural ideas, local receptive fields, shared weights, and spatial or temporal subsampling, are responsible for the power of CNN. Usually,

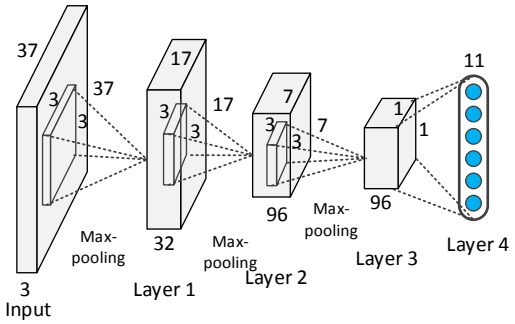

Figure 2: The structure of color CNN.

CNN-based deep feature learning pipeline has two stages. The first is to learn a CNN model in a supervised manner. The second is to extract deep feature from last several layers of CNN. Following this pipeline, we extract two deep features to jointly describe the object from the view of structure and color.

2.1 Deep Structure Feature

It has been verified that CNN model supervised by object categories can not only abstract rich semantic information but also capture local structure property of the object [20]. Since the 1000 object categories in ILSVRC12 dataset include person and vehicle, which are the most often retrieved objects in surveillance scenarios, we choose ILSVRC12 dataset to train a CNN model to extract rich semantic structure information. Specifically, we adopt the exact CNN architecture specified in [20] and train the model with the help of Caffe [6]. Considering higher layers in CNN models generally produce more discriminative features, we extract deep structure feature from the penultimate layer. During the image preprocessing in feature extraction phase, we first resize the input image to 225×225, then subtract the mean activity over the training set from each pixel.

2.2 Deep Color Feature

Although deep structure feature delivers rich semantic structure information, it does not contain enough color information, which is rather crucial in describing small ob-

jects. Therefore, inspired by [17], we build a CNN model to describe the object from the view of color. Based on the linguistic study of Berlin and Kay [2], we select 11 basic color terms of the English language: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow, as the supervisory information when training CNN on Google color name set. By training CNN model to classify 11 color names, we can learn deep color representation which is robust to illumination change and color distortion across different scenes. The detailed architecture of the CNN model is shown in Figure 2. To comprehensively describe the spatial color distribution, we first preprocess the input image to be of 55×55 size and zero mean value, then divide it averagely into 16 overlapping 37×37 patches and send them into color CNN to get 16 96-d features, which are concatenated into 1536-d deep color feature. To test the discriminability of the deep color feature, we perform a classification experiment on Ebay dataset [18] and achieve an accuracy of 0.80, which is higher than that obtained by using traditional hand-crafted color features, such as color histogram.

2.3 Multi-view Deep Features Encoding

The above multi-view deep features are rather descriptive but not compact enough. For searching objects in large scale dataset, high dimensionality will significantly corrupt the retrieval efficiency. Due to the efficiency in both storage and speed, hashing based approximate nearest neighbor search methods have attracted much attention in the past years. Hashing converts all feature representation of images into binary codes and then conducts a bitwise XOR operation in very fast speed. LSH [4] is one of the most popular hashing algorithms. Using LSH, we encode each deep feature into one 256-bit binary code as a compact descriptor for visual object. The time cost is smaller for computing Hamming distance than Euclidean distance. Since in modern CPUs, it only takes one CPU cycle to compute Hamming distance for 128-bit hash code in Hamming space. To further improve retrieval accuracy, we combine the two compact descriptors with a late fusion strategy to obtain final retrieval results.

3. EXPERIMENTS

Since there is no available standard large scale database in the field of object retrieval in surveillance scenarios, we collect surveillance videos with HD cameras mounted at residential entrances inside university. We adopt background subtraction and object tracking [5] to extract moving objects, and build a dataset consisting of 100k objects. The dataset covers various weather conditions, light changes, poses, viewpoints, and partial occlusions. A total of 300 query objects, 100 vehicles and 200 persons, are selected to evaluate the retrieval performance. Mean average precision (MAP) is used to measure the retrieval performance of different methods.

3.1 Comparison with Hand-crafted Features

To show the effectiveness of the multi-view deep features, we compare them with some hand-crafted visual descriptors. The comparison results are demonstrated in Table 1. We calculate the MAP of top 10k returned objects for person and vehicle respectively. Six hand-crafted visual descriptors are used as comparison features. "ColorHist" is a 81-d color histogram in HSV space, while "ColorMoment [14]" is a 9-d feature obtained by calculating first, second and third moments of image pixels in LAB space. "GLCM [11]" is a 48-d texture feature based on gray-level co-occurrence matrix and "WaveletTransform" is a 20-d texture feature based on wavelet transform. "CannyEdgeHist" is a 64-d feature describing the edge information returned by canny edge detector, while "ShapeContext [1]" is a 72-d feature describing the local shape information of objects. It can be found that color descriptors are more useful than texture, edge and shape descriptors for small object retrieval.

Table 1: Comparison with hand-crafted features.

	MAP(person)	MAP(vehicle)
WaveletTransform	0.1801	0.1750
ShapeContext [1]	0.1978	0.1921
CannyEdgeHist	0.2119	0.1835
GLCM [11]	0.2390	0.2119
ColorMoment [14]	0.2825	0.2998
ColorHist	0.2876	0.3648
colCNN-LSH	0.4378	0.4113
strCNN-LSH	0.4729	0.3991
colCNN	0.5007	0.4609
strCNN	0.6122	0.4321
str+colCNN-LSH	0.5955	0.4732
str+colCNN	**0.6585**	**0.4893**

Table 2: Retrieval cost of different methods.

	Retrieval cost(s)
colCNN-LSH	0.935
strCNN-LSH	0.971
str+colCNN-LSH	1.243
ColorMoment [14]	1.861
WaveletTransform	1.871
GLCM	1.936
CannyEdgeHist	1.954
ColorHist	1.964
ShapeContext [1]	1.979
colCNN	5.758
strCNN	12.445
str+colCNN	18.109

For deep features extracted from CNN, "strCNN" is the 4096-d deep structure feature and "colCNN" is the 1536-d deep color feature, which describe the object from the view of structure and color respectively. "str+colCNN" indicates the late fusion of "strCNN" and "colCNN", which performs the best both in person and vehicle retrieval with a relative improvement of 7.56% and 6.16% in MAP than the best results achieved by single descriptor. "strCNN-LSH" and "colCNN-LSH" is the 256-bit compact feature encoded from "strCNN" and "colCNN" with LSH respectively. Although they achieve lower MAP than "strCNN" and "colCNN", they perform better than all of the remaining comparison features and save retrieval time greatly. By fusing "strCNN-LSH" and "colCNN-LSH", we get "str+colCNN-LSH", which reduces the retrieval time from about 18 seconds to 1 seconds, with only a little decrease in MAP compared with "str+colCNN".

3.2 Comparison of Retrieval Cost

As shown in Table 2, six hand-crafted features all cost less than 2 seconds in retrieval time due to the low dimensionality. And it is no surprise that the 4096-d deep structure feature costs the most retrieval time, which is then greatly

Figure 3: Comparison of the ranking examples.

reduced by encoding deep features into 256-bit binary codes with LSH. All the retrieval experiments are conducted on Intel i5-2400 CPU 3.1 GHz with 20 GB memory. Figure 3 shows some retrieval examples of our proposed approach as well as several comparison methods.

4. CONCLUSIONS

In this paper, we propose an effective multi-view deep features learning approach for small object retrieval in surveillance scenarios. For structure view, we train a CNN model for object category classification to extract deep structure feature. For color view, we propose a CNN model for color name classification to extract deep color feature. Then we use LSH to encode deep features into short binary codes to accelerate retrieval efficiency and fuse the compact deep features to increase retrieval accuracy. Compared with several popular visual descriptors, our proposed approach achieves the best performance.

5. ACKNOWLEDGMENTS

This work was supported by 863 Program 2014AA015104, and National Natural Science Foundation of China 61273034, and 61332016.

6. REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *TPAMI*, 24(4):509–522, 2002.

[2] B. Berlin. *Basic color terms: Their universality and evolution*. Univ of California Press, 1991.

[3] S. Calderara, R. Cucchiara, and A. Prati. Multimedia surveillance: content-based retrieval with multicamera people tracking. In *IWVSSN*, pages 95–100. ACM, 2006.

[4] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Symposium on Theory of computing*, pages 380–388. ACM, 2002.

[5] S. Chris and G. W. E. L. Adaptive background mixture models for real-time tracking. In *CVPR*, volume 2. IEEE, 1999.

[6] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[7] A. Krizhevsky and G. Hinton. Using very deep autoencoders for content-based image retrieval. In *ESANN*. Citeseer, 2011.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[9] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *NC*, 1(4):541–551, 1989.

[10] D. Lowe. Object recognition from local scale-invariant features. In *CV*, volume 2, pages 1150–1157. Ieee, 1999.

[11] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *PAMI*, 18(8):837–842, 1996.

[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[13] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[14] S. M. Singh and K. Hemachandran. Image retrieval based on color moments. *Digital Image Processing*, 4(16):910–916, 2012.

[15] K. Teng, J. Wang, M. Xu, and H. Lu. Mask assisted object coding with deep learning for object retrieval in surveillance videos. In *MM*, pages 1109–1112. ACM, 2014.

[16] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen. Person attribute search for large-area video surveillance. In *HST*, pages 55–61. IEEE, 2011.

[17] Y. Wang, J. Liu, J. Wang, Y. Li, and H. Lu. Color names learning using convolutional neural networks. In *ICIP*. IEEE, 2015.

[18] J. V. D. Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *CVPR*, pages 1–8. IEEE, 2007.

[19] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In *ICIP*, pages 2937–2940. IEEE, 2011.

[20] M. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. *arXiv preprint arXiv:1311.2901*, 2013.