Mask Assisted Object Coding with Deep Learning for Object Retrieval in Surveillance Videos

Kezhen Teng[1], Jinqiao Wang[1], Min Xu[2], and Hanqing Lu[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100190
[2]iNEXT, School of Computing and Communications, University of Technology, Sydney, Australia
{kezhen.teng,jqwang,luhq}@nlpr.ia.ac.cn, min.xu25@gmail.com

ABSTRACT

Retrieving visual object from a large-scale video dataset is one of multimedia research focuses but a challenging task due to imprecise object extraction and partial occlusion. This paper presents a novel approach to efficiently encode and retrieve visual objects, which addresses some practical complications in surveillance videos. Specifically, we take advantage of the mask information to assist object representation, and develop an encoding method by utilizing highly nonlinear mapping with a deep neural network. Furthermore, we add some occluded noise into the learning process to enhance the robustness of dealing with background noise and partial occlusions. A real-life surveillance video data containing over 10 million objects are built to evaluate the proposed approach. Experimental results show our approach significantly outperforms state-of-the-art solutions for object retrieval in large-scale video dataset.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Theory

Keywords

Object retrieval, Video search, Autoencoder

1. INTRODUCTION

The increasing number of cameras produce a huge amount of video data. It is an urgent need to develop intelligent techniques for object indexing and retrieval in a large-scale video data. Content-based object retrieval is an advanced multimedia application, which is benefitted from efficient representation [8] and indexation [13, 7] approaches. Although this problem attracts increasing research interests [14, 4, 10, 1, 7] in recent years, developing an efficient object retrieval solution, especially for surveillance videos, is still a challenging task, primarily due to three issues. First, such a solution involves a series of processing, such as background modelling, object extraction and representation, which are all crucial but challenge tasks. Second, the appearance of the same object is inconsistent under different cameras, which add the difficult with the combination of view and pose change. Finally, complex conditions such as low resolution, illumination, noise, shadows, and partial occlusions embodied with surveillance environment, make the objects inherently subtle and even vague for human to recognize, thus considerable computations are required to discern them.

As analysis above, in this paper a content-based object retrieval approach is proposed based on deep learning, which is to encode object together with mask information efficiently. A multi-model deep learning strategy is proposed to restructure the training set and improve retrieval performance when the background is very noisy and with partial occlusions. Moreover, we collect a real-life dataset contains more than 10 million objects for evaluation. Our method outperforms other hash methods 10% on object retrieval in surveillance video.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654981 .

2. RELATED WORK

Object retrieval in surveillance mainly focuses on the objects, such as persons and vehicles [5, 18, 15]. Calderara *et al.* [2] proposed the architecture for person retrieval in multi-camera surveillance with the non-overlapping views, and estimated the color probability distribution with a mixture of Gaussians. Feris *et al.* [5] extracted a set of fine-grained attributes for moving vehicles such as time, direction, dominant color, dimensions and speed, etc. Then they automatically ingested the attribute metadata into a back-end database system through a web-based service-oriented architecture. Thornton *et al.* [15] focused on attributes of gender, hair/hat color, clothing color, and bag (if any) position and color. A generative model was proposed to build the corresponding descriptors for person search. Yang and Yu [18] combined color histograms and three different texture descriptors to real-time recognize eight kinds of clothes in surveillance videos.

Recent researches indicate that the generic descriptors extracted from the deep neural networks are very powerful. Hinton [9] applied a 19 layers very deep autoencoder to encode natural images into binary codes and achieve satisfactory results. However, it focused on the whole image, not the object we are interested in. Inspired by [9], in this pa-

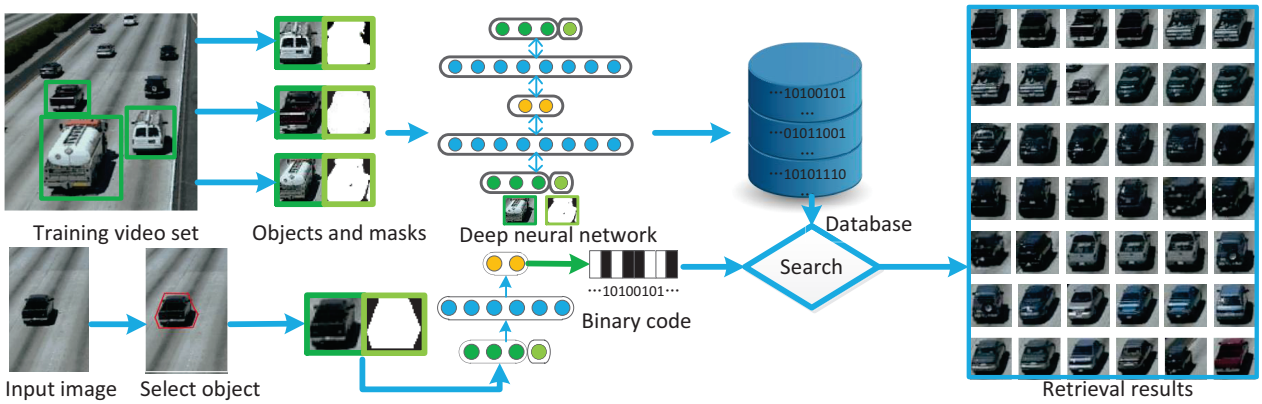

Figure 1: Overall framework.

per through mapping the object and mask information extracted from surveillance videos to short binary codes with deep learning, a very deep representation is learned to ensure similar objects have similar binary codes.

3. OVERVIEW

The framework of the proposed approach is illustrated in Fig. 1. In the training phase, background subtraction and multi-object tracking [19] is performed for surveillance videos to extract active objects such as a person and a car. Two sub-images are extracted for each object: the color image of the object block and the corresponding mask image. Then a multimodel deep neural network is trained to encode the object and mask image into a 128 bit binary vector. In the retrieval phase, user can select an interesting object from the query image through an interactive operation. Afterwards the selected object is encoded with the learned deep neural network. Then the generated 128 bit code is used to retrieve similar objects in the database with hamming distance.

4. DEEP OBJECT REPRESENTATION

Autoencoder (AE) was firstly proposed by Hinton in 2002. As a deep neural network (DNN), it is composed with stacks of restricted Boltzmann machine (RBM) and is used to transfer input vectors into relative short codes with a highly nonlinear mapping function. With the progress of DNN theory and GPU technique, it becomes practical to train AE with acceptable computation ability. The training process is divided into two phases: unsupervised pre-train and supervised fine tune. To initialize the deep neural network, the encode part is trained generatively layer by layer. Then the weights and off-set vectors are unrolled to initialize the decode part. Afterwards, the whole neural network is fine-tuned with back propagation.

4.1 Mask assisted object coding

While in surveillance videos, the same object may appear in different scenes, or the same location in a frame but at different time. Thus they are surrounded with different background noise. Sometimes, e.g. in the case of traffic jam, an object is surrounded by many other objects. The denoised autoencoder (DAE) [16] is designed to increase robustness when dealing with global noise. It is trained to reconstruct a clean "repaired" input from its noisy version. Formally, the input v's noised version $\tilde{v}$ is construct through a stochastic mapping. The noisy version $\tilde{v}$ will be then mapped through AE to reconstruct a clean version of $\tilde{v}$ by $\hat{v}$. Note that the reconstruction error is $L = \sum_{i=1}^{N} |v_i - \hat{v}_i|^2$.

However, DAE is not proper for our object retrieval problem in surveillance videos due to two reasons. First, the encoding phase of DAE is encouraged to be robust for noise in the whole image, not the background area of image. Second, the distribution of foreground object is related to background area. If noise is only added to the background area, the training process will be confused. Details are shown as follows.

Suppose the foreground and background part of input image v are represented by f and b respectively. The appearance of foreground and background are represented by α and β respectively. And objective factors such as lighting, color deviation and stochastic noise are governed by θ. Then the distribution of v is formulated as,

$$p(v|f,b,\alpha,\beta,\theta) \propto p(v|f,b)p(f|\alpha,\theta)p(b|\beta,\theta) \tag{1}$$

Eq. 1 demonstrates that foreground and background are both related to θ. If the noise is added to the background area of image, the right part of Eq. 1 will be changed to $p(v|f,b)p(f|\alpha,\theta)p(b|\beta,\tilde{\theta})$, this will confuse the learning process.

Since moving objects are usually extracted with background subtraction, it is natural to incorporate the mask information to enhance object representation. Inspired by [11, 12], to learn a feature representation robust to different background images, we adopt a mask assisted multimodal deep neural network to learn discriminative object codes.

The mask image is represented by m, which is as the object context information. Then the object encoding solution α can be solved with autoencoder as,

$$\begin{aligned} arg\max_{\alpha} \quad & p(\hat{v},\hat{m}|v,m) \\ \propto \quad & p(\hat{v},\hat{m}|b,f)p(b,f|v,m) \\ \propto \quad & p(\hat{v},\hat{m}|b,f)p(f|\alpha,\theta)p(b|\beta,\theta)p(\alpha,\beta,\theta|v,m) \end{aligned} \tag{2}$$

In fact the optimization problem is often recast as,

$$arg\min_{\alpha} \quad L = |v-\hat{v}|^2 + |m-\hat{m}|^2 \tag{3}$$

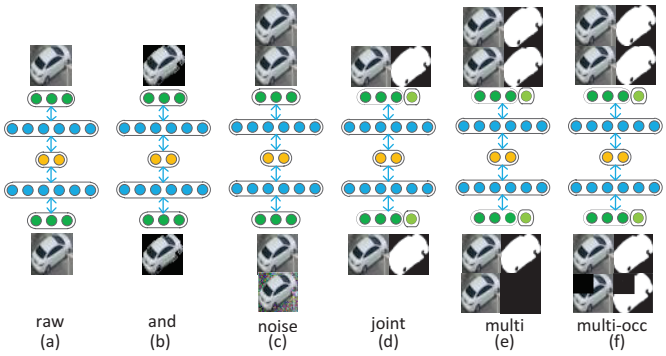

Figure 2: Different coding designs with DNN.

By introducing mask information, the multimodel autoencoder could learn the context appearance of object by minimizing the reconstruction error.

To illustrate our multimodel learning structure, different coding designs of deep neural networks are given in Fig. 2. (a) is to directly learn the feature codes with autoencoder. (b) is to learn with the object region only. (c) is to learn with denoising autoencoder. (d), (e) and (f) are our multimodel solutions. (d) is to jointly learn with object images and mask images together. (e) is to multimodel learn with object images and mask images. For multimodel neural network, the training set is copied and divided into two totally equal parts. The first part includes input object images and mask images. In the second part, mask images are all transformed to pure black images. In the pre-training phase, the two parts are both used. while in the fine-tuning phase, both parts are tuned by the first part. Thus even for the user input without effective mask image, this multimodel autoencoder is expected to reconstruct object image with complete mask image. In this way, the deep neural network is supposed not only to encode the object itself, but also to segment it from different background.

4.2 Occlusion processing

Another challenge confronted by surveillance videos is the partial occlusion problem, especially in crowed scene such as shopping mall and residential quarters. To address this challenge, we further add some occlusion noise into the multimodel autoencoder, as shown in Fig. 2(f). The training set is designed to comprise occlusion noise. Specifically, we insert some random patches into the object image, and for each patch the pixels are set to zero. Then these noised images and un-noised images compose the whole training set. In the pre-training phase, the noised images and the un-noised images are both used. While in the fine-tuning phase, those noised patch is fine-tuned with the un-noised patch. In this way, the learned object codes could reconstruct the missing regions for partial occluded object.

5. EXPERIMENTS

5.1 Dataset and setting

Since there no available large scale surveillance dataset that provides the object and mask images, to evaluate the performance of the proposed approaches we collect surveillance videos with HD cameras mounted at residential entrances inside university. We utilize background subtraction and object tracking to extract object images and corresponding mask images, and build a dataset about 10 million objects. The object dataset covers different weather conditions, different lighting effects, and different periods of time.

The color object blocks are resized to 32×32 with three channels and the binary mask images are resized to 32×32 with only one channel. Then a 4096 dimension vector is built to represent the object. To obtain more training instances, all images are flipped left to right to double the training set.

We train on 10 million images that have been preprocessed by subtracting from each pixel to its mean value over all images and then dividing by the standard deviation of each pixels over all images. The first RBM in the stack has 8192 binary hidden units and 4096 linear visible units with unit variance Gaussian noise. All the remaining RBM's have N binary hidden units and $4N$ binary visible units until it reaches the designed size. Details of training process is similar to [9]. The entire training procedure for each multimodel autoencoder takes about 1 day on a Nvidia Tesla M2075.

Table 1: Retrieval results of different approaches with 128 bit codes (MAP %).

	original	noise	occlude(0.25)	occlude(0.5)
ITQ(32bit)	39.09	38.28	15.45	4.93
raw(32bit)	46.93	37.92	18.80	5.21
ITQ	52.43	50.20	39.11	6.08
LSH	49.41	46.45	35.81	6.75
SH	48.22	44.98	30.64	8.19
raw	50.02	46.81	23.57	5.90
and	52.98	52.98	40.26	11.39
noise	53.85	50.10	22.64	11.85
joint	62.11	60.91	39.71	12.33
multi	64.91	65.46	40.38	12.56
multi-occ	65.02	60.25	41.27	18.61

5.2 Retrieval results

Total 43 queries are used to evaluate the proposed approach. Mean Average Precision (MAP) is used to quantitatively measure the performance. To compare the robustness for background noise and partial occlusion, four groups of queries images were carried out for each approach. The first group includes clean object images and mask images. In the second group, Gaussian noise was added to query images. In the third group, 0.25% of object region in queries is random occluded. The same setting was used for the fourth group queries with occluded block size of 50% of query objects. In addition to the original autoencoder, some classic hashing approaches LSH (locality sensitive hashing) [3], SH (spectral hashing) [17], ITQ (iterative quantization) [6] are also compared, and retrieval results of different approaches with 128 bit codes are shown in Tab. 1.

From Tab. 1, "raw" that directly learns with autoencoder [9] achieves satisfactory results even for short codes (32 bit). But it seems that those linear hash methods achieve better robustness in partial occlusion conditions for 128 bit codes. When mask image is introduced, the retrieval performance is effectively boost even for one modality learning. With 128 bit codes, the MAP is increased 12.09% for "joint" with simple object images and mask images than "raw". For the multi-modality learning "multi", as our expect, the MAP

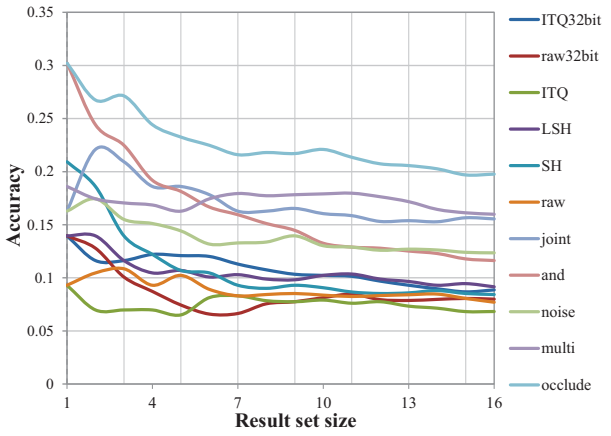

Figure 3: Retrieval accuracy of different approaches.

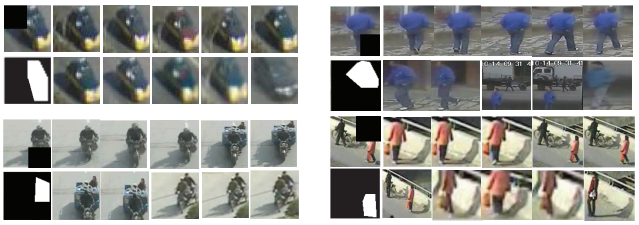

Figure 4: Retrieval results with occlusion.

raised 4.55% in noised conditions. And for the occluded learning "multi-occ", the MAP increased 6.28% for half occlusions than "multi", 10.42% than SH [17].

Fig. 3 shows the accuracy with different result set size in the experiment with 50% occlusion. The mask assisted object coding approaches "joint', "multi" and "multi-occ" outperform hash methods such as LSH [3], SH [17] and ITQ [6]. And the "multi-occ" reaches the best result, which shows the effective of our multimodel DNN for robustness of occlusion. Fig. 4 gives some examples in occlusion conditions.

6. CONCLUSION

In this paper, a mask assisted object encoding approach is presented to boost retrieval performance in surveillance videos. The object image and mask image are used to learn multimodel deep neural network that map similar objects to similar binary codes. Additionally, occlusion noise is added into the training set for reconstructing the whole object to handle partial occlusion. Experiments and comparisons demonstrate the efficiency of the proposed approach.

7. ACKNOWLEDGEMENT

This work was supported by 973 Program (2010CB327905), 863 Program (2014AA015104 and 2014AA015105), and National Natural Science Foundation of China (61273034, and 61332016).

8. REFERENCES

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2911–2918, 2012.

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

[3] M. Charikar. Similarity estimation techniques from rounding algorithm. In *ACM symposium on Theory of computing*, pages 380–388, 2002.

[4] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 253–262, 2005.

[5] R. Feris, B. Siddiquie, Y. Zhai, J. Petterson, L. Brown, and S. Pankanti. Attribute-based vehicle search in crowded surveillance videos. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 18, 2011.

[6] Y. Gong and S. Lazebnik. Iterative quantization: a procrustean approach to learning binary codes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[7] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2916–2929, 2013.

[8] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311, 2010.

[9] A. Krizhevsky and G. E. Hinton. Using very deep autoencoders for content-based image retrieval. In *European Symposium on Artificial Neural Networks*, 2011.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178, 2006.

[11] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2480–2487, 2012.

[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

[13] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168, 2006.

[14] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477, 2003.

[15] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen. Person attribute search for large-area video surveillance. In *IEEE Int. Conf. on Technologies for Homeland Security (HST)*, pages 55–61, 2011.

[16] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[17] Y. Weiss., A. Torrlaba, and R. Fergus. Spectral hashing. In *Advances in Neural Information Processing System (NIPS)*, 2008.

[18] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In *18th IEEE International Conference on Image Processing (ICIP)*, pages 2937–2940, 2011.

[19] Y. Zhang, J. Wang, W. Fu, H. Lu, and H. Xu. Specific vehicle detection and tracking in road environment. In *ICIMCS*, pages 182–186, 2011.