Video Semantic Indexing using Object Detection-Derived Features

Kotaro Kikuchi, Kazuya Ueki, Tetsuji Ogawa, and Tetsunori Kobayashi
Dept. of Computer Science, Waseda University, Japan

***Abstract*—A new feature extraction method based on object detection to achieve accurate and robust semantic indexing of videos is proposed. Local features (e.g., SIFT and HOG) and convolutional neural network (CNN)-derived features, which have been used in semantic indexing, in general are extracted from the entire image and do not explicitly represent the information of meaningful objects that contributes to the determination of semantic categories. In this case, the background region, which does not contain the meaningful objects, is unduly considered, exerting a harmful effect on the indexing performance. In the present study, an attempt was made to suppress the undesirable effects derived from the redundant background information by incorporating object detection technology into semantic indexing. In the proposed method, a combination of the meaningful objects detected in the video frame image is represented as a feature vector for verification of semantic categories. Experimental comparisons demonstrate that the proposed method facilitates the TRECVID semantic indexing task.**

1. Introduction

Since vast amounts of video data have been uploaded on the Internet, efficient search technologies for videos are being explored. In fact, many attempts have been made to assign a semantic tag to a video clip (referred to as *semantic indexing*) [1], [2], [3], making it possible to achieve video retrieval without depending on meta data, such as titles and descriptions.

Let us consider as an example the semantic tag "Bicycling." The corresponding video clip includes a scene of a person riding a bicycle. It is assumed that the frames of this video contain these two key objects, a person and a bicycle. The type of road on which the bicycle is being ridden does not, however, characterize "Bicycling" to a very great extent. This indicates that the video frame contains not only meaningful objects but also a background region that can be ignored when determining the semantic tag, and the former should be emphasized in feature extraction for semantic indexing.

In relevant previous work for TRECVID [4], [5], which is an international competition on video retrieval, various features, such as local features, motion features, and acoustic features, were applied and combined [1]. In a more sophisticated manner, intermediate layer outputs of well-trained CNN were extracted and then taken as the inputs of support vector machines (SVMs), yielding a high accuracy on TRECVID 2015 data [2], [3]. It should be noted that these methods perform semantic category verification by using features extracted from all the pixels over a video frame image. In this case, the background information used could adversely affect the performance of the semantic tag indexing.

In contrast, we define the meaningful objects, the combination of which can characterize semantic categories (e.g., bicycle, person), and attempt to explicitly use the probabilities of these objects existing in the video frame as feature representations for semantic indexing. These object detection-derived features are extracted by using faster region-based CNN (fRCNN) [6], which explicitly considers object detection inside the network, instead of CNN [3]. In this study, a system using the proposed fRCNN-derived features was integrated with that using conventional CNN-derived features [3]. The developed system can be robust against the harmful effects derived from the meaningless background. In addition, the proposed feature can be of use in particular in the case of a small amount of training data, because the co-occurrence of the objects, which contributes to semantic indexing performance, is directly represented.

The rest of the present paper is organized as follows. In Section 2, the proposed feature extraction method exploiting object detection is described. In Section 3, experimental comparisons are presented to demonstrate the effectiveness of the proposed feature for the TRECVID data. In Section 4, the present paper is concluded and some future work described.

2. Object detection-derived features for semantic indexing

The study presented in this paper focused on CNN-based feature extraction in a CNN/SVM tandem connectionist architecture, which is the core system used and was shown to be effective in the semantic indexing task of TRECVID 2015 [3]. This core system and conventional CNN-derived features [3], which handle the entire image, are briefly explained in 2.1 and 2.2, respectively. Feature extraction using fRCNN-based object detection is proposed in 2.3, yielding the fRCNN/SVM tandem system. In addition, an attempt to fuse the systems using these two features is described in 2.4.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

2.1. CNN/SVM tandem connectionist architecture

We exploited the CNN/SVM tandem connectionist architecture for semantic indexing in which intermediate layer outputs or estimated targets in well-trained CNN are extracted and then taken as inputs of an SVM-based verifier of semantic categories. In this architecture, a transfer learning concept is exploited to improve the robustness of semantic indexing systems against low resources for training. The CNN is developed on large-scale ImageNet data [7], which have primitive objects in common with those of the target task (i.e., TRECVID semantic indexing), and the SVM is trained on task-specific limited data. It should be noted that in most cases the number of positive samples for each category is very limited in TRECVID data: there are only several hundred or sometimes less than one hundred samples for each category. The CNN/SVM tandem architecture based on transfer learning is therefore a better choice than a simple architecture using only a CNN trained from scratch on very limited TRECVID data; the former architecture makes it possible to improve the reliability of the total system by using reliable CNN-derived features, while the latter system tends to yield unreliable results.

2.2. CNN-derived features

After the breakthrough in image classification in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, CNNs have been used as one of the very effective feature extraction methods, not only in image classification but also in other tasks [8]. We also use a CNN, specifically AlexNet [9], as a feature extractor. AlexNet contains five convolutional layers and three fully-connected layers. The network used is trained on the ImageNet database, which includes 1.2 million images and 1,000 categories, and is available in the Caffe library [10]. After inputting an image to the AlexNet, we extract a 4,096-dimensional vector from the seventh layer and use it as a visual feature. This approach has been shown to be very effective also in video semantic indexing [3].

After extracting 4,096-dimensional vectors, we use SVMs with a radial basis function (RBF) kernel to train a model for each semantic category.

2.3. fRCNN-derived features

Since CNN-derived features include irrelevant background information, object-oriented features are expected to be complementary. Thus, in addition to CNN-derived features, we extract object detection-derived features and compose a vector representation from a combination of detected objects. We chose fRCNN [6] to detect specific objects in an image. fRCNN is a fast and high-performance object detection method based on CNNs, which includes a region proposal network (RPN) and an object detection network. It achieved a state-of-the-art performance on the PASCAL VOC benchmark dataset. An overview of feature extraction in the fRCNN/SVM tandem system is illustrated in Fig. 1. When we input an image $\mathbf{x}$ to fRCNN, we can obtain approximately 200 bounding boxes and their probability scores $p_{ji} = p(j|\mathbf{x}, i)$ for individual object categories, where i denotes an index of a bounding box and j denotes an index of an object category.

We select the maximum probability output of each category over all the bounding boxes and concatenate them to create a feature vector $\mathbf{o}$ as

$$\hat{p}_j = \underset{i}{\operatorname{argmax}}\, p_{ji}, \tag{1}$$

$$\mathbf{o} = [\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_C], \tag{2}$$

where C denotes the number of target categories in object detection ($C = 20$ in our experiment).

After extracting C-dimensional feature vectors, we train SVMs in the same manner as in 2.2.

2.4. System integration

The fRCNN/SVM tandem system is integrated with the CNN/SVM system as

$$s(\mathbf{x}) = \alpha f(\mathbf{h}; \theta_{\text{cnn/svm}}) + (1 - \alpha) g(\mathbf{o}; \theta_{\text{frcnn/svm}}), \tag{3}$$

where $f(\mathbf{h}; \theta_{\text{cnn/svm}})$ and $g(\mathbf{o}; \theta_{\text{frcnn/svm}})$ denote a score yielded by the CNN/SVM system and the fRCNN/SVM system, respectively, and $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{o}$ denote the input keyframe image, its intermediate layer outputs through CNN, and its targets estimated through fRCNN, respectively. The fusion weight α is arbitrarily determined to be a value from zero to one.

3. Semantic indexing experiment

Experimental comparisons were conducted in semantic indexing to verify the effectiveness of the proposed method. The semantic indexing systems evaluated were as follows.

- **CNN**: CNN/SVM tandem system, which performs SVM-based verification using CNN-derived features [3]
- **fRCNN (Proposed)**: fRCNN/SVM tandem system, which performs SVM-based verification using fRCNN-derived features
- **CNN+fRCNN (Proposed)**: score-level system integration of CNN/SVM and fRCNN/SVM systems

3.1. Video materials and tasks

This subsection describes the data and evaluation procedure on TRECVID.

3.1.1. TRECVID2010 dataset. The developed semantic indexing systems were evaluated on the TRECVID2010 dataset [11]. These data consist of Internet archive videos with creative commons licenses. To reduce the ambiguity of the videos in terms of the semantic tags assigned, each video was segmented into short video clips, referred to as "shots."

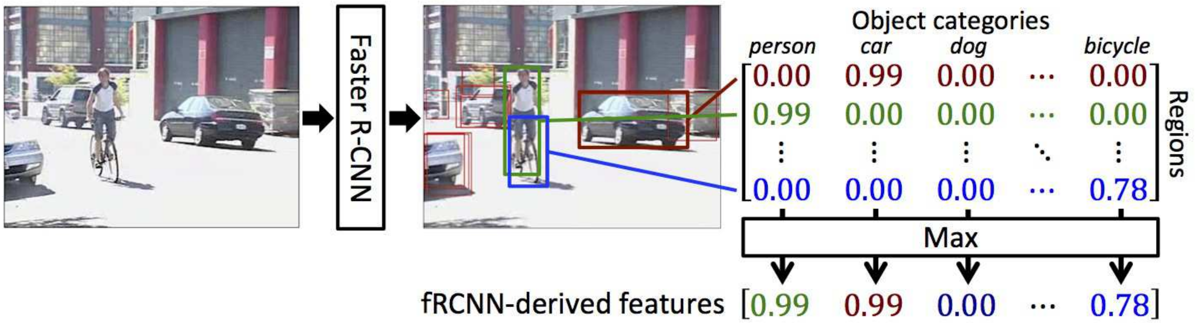

Figure 1. Overview of fRCNN-based feature extraction.

The TRECVID2010 dataset includes 119,685 training shots (approximately 200 hours) and 144,988 testing shots (approximately 200 hours). In addition, a "keyframe," which is the single video frame image selected from the shot, is used for semantic indexing. Semantic "concepts," which correspond to labels of semantic classes, include objects, events, and scenes, each of which has different properties. The concept labels are provided using the collaborative annotation scheme [12], [13]. In the task of semantic indexing at the TRECVID2010 workshop, participants were required to detect 50 semantic concepts.

3.1.2. Semantic indexing task. Semantic indexing is designed as concept verification, i.e., it determines whether the target concept should be assigned to an input shot or not. This verification system can be implemented with a two-class classifier developed for each concept.

For testing, we used the evaluation criterion for semantic indexing of TRECVID2010, namely, the average precision (AP). The AP of each category is defined as

$$AP = \frac{1}{N_{\mathrm{pos}}^{(\mathrm{te})}} \sum_{r=1}^{N^{(\mathrm{te})}} P_r \cdot Rel_r, \tag{4}$$

where $N^{(\mathrm{te})}$ denotes the number of test shots, $N_{\mathrm{pos}}^{(\mathrm{te})}$, the number of positive test shots, and r, the rank in the ordered list of results retrieved from $N^{(\mathrm{te})}$ shots. P_r is defined as the precision computed at the r-th rank and Rel_r takes the value 1 or 0, representing relevant or irrelevant, respectively. Finally, the developed systems were evaluated using the mean AP (mAP): the AP scores averaged across all concepts.

At TRECVID2010, the participants evaluated the entire testing set (144,988 shots), outputted their scores for each concept, and submitted ranked lists of the top 2,000 shots for each of 50 concepts to calculate the mAP.

3.2. Experimental setups

For training the SVMs, we collected positive samples using collaborative annotation [12], [13]. Each concept had 10 to 3,000 positive shots, respectively. Negative shots were randomly selected from non-positive shots. The number of negative examples was adjusted such that the number of positive and negative shots would be 30,000 in total.

TABLE 1. TWENTY TARGET OBJECTS TO BE DETECTED IN PASCAL VOC 2007 DATASET

aeroplane	bicycle	bird	boat	bottle
bus	car	cat	chair	cow
diningtable	dog	horse	motorbike	person
pottedplant	sheep	sofa	train	tvmonitor

Twenty kinds of target object, as shown in Table 1, were detected using fRCNN. We manually selected 30 out of 50 concepts that were related to the target objects, as listed in Table 2, because our aim is to clarify the effect of object detection on feature extraction for semantic indexing.

Since the number of target objects was 20, fRCNN-derived features were represented by 20-dimensional vectors. The VGG-16 model, which is a very deep neural network with 16 layers [14], was exploited inside fRCNN. We used a pre-trained model provided on GitHub [15]. This model was trained on the PASCAL VOC 2007 dataset.

For the system fusion, the fusion weight α described in Eq. 3 was empirically set to 0.5.

3.3. Experimental results

Table 3 shows the mAPs and APs for individual concepts, obtained from the CNN/SVM system, fRCNN/SVM system, and their integration. The CNN/SVM, fRCNN/SVM, and CNN/SVM+fRCNN/SVM systems gave an mAP of 12.48, 2.61, and 12.78, respectively. The developed CNN/SVM+fRCNN/SVM system yielded improvements as compared to the conventional CNN/SVM system for 20 of all the 30 concepts. Figures 2 and 3 describe the keyframe images of the top 20 shots estimated as the concept "Ground_Vehicles" by using the CNN/SVM and fRCNN/SVM systems, respectively. The concept "Ground_Vehicles" is defined on TRECVID as ground vehicles, such as automobiles, bicycles, buses, motorbikes, autotrucks, and so on. These figures indicate that the CNN/SVM system tends to rank higher the images of automobiles, being large and located at the center, while the fRCNN/SVM system can detect not only automobiles but

TABLE 2. SELECTED 30 SEMANTIC CONCEPTS IN TRECVID 2010 DATASET

Adult	Airplane_Flying	Animal	Asian_People	Bicycling
Boat_Ship	Bus	Car_Racing	Cheering	Computer_Or_Television_Screens
Dark_skinned_People	Demonstration_Or_Protest	Female_Person	Female_Human_Face_Closeup	Ground_Vehicles
Hand	Infants	Instrumental_Musician	Male_Person	News_Studio
Old_People	Plant	Running	Singing	Sitting_Down
Swimming	Throwing	Vehicle	Walking	Walking_Running

TABLE 3. AVERAGE PRECISION FOR CNN/SVM SYSTEM, fRCNN/SVM SYSTEM, AND THEIR INTEGRATION

Concept	CNN	fRCNN	CNN+fRCNN
Adult	5.58	3.83	**5.68**
Airplane_Flying	15.02	3.61	**16.10**
Animal	16.50	11.37	**19.65**
Asian_People	0.57	0.00	**0.62**
Bicycling	11.49	1.62	**12.24**
Boat_Ship	**23.51**	3.90	22.86
Bus	2.09	2.09	**2.58**
Car_Racing	**5.35**	0.03	3.31
Cheering	2.35	0.00	**2.41**
Computer_Or_Television_Screens	5.31	3.17	**5.77**
Dark-skinned_People	5.81	0.00	**5.82**
Demonstration_Or_Protest	9.02	0.10	**9.11**
Female_Person	**10.73**	0.08	10.68
Female-Human-Face-Closeup	**14.26**	0.00	**14.26**
Ground_Vehicles	21.83	20.73	**26.69**
Hand	14.01	0.15	**14.21**
Infants	3.73	0.00	**3.97**
Instrumental_Musician	25.16	0.46	**25.63**
Male_Person	**5.88**	2.67	5.45
News_Studio	**67.21**	0.35	66.91
Old_People	3.54	0.34	**3.62**
Plant	13.17	4.54	**13.81**
Running	3.33	0.08	**3.40**
Singing	6.97	0.05	**6.98**
Sitting_Down	**0.08**	0.00	**0.08**
Swimming	**37.19**	0.67	36.31
Throwing	**5.79**	0.01	5.53
Vehicle	24.67	15.43	**26.37**
Walking	**7.64**	1.50	7.40
Walking_Running	**6.55**	1.64	5.85
Mean Average Precision	12.48	2.61	**12.78**

also buses and bicycles as target objects. In addition, the fRCNN/SVM system successfully detected objects that are very small and located near the edge of the image. The use of fRCNN-derived features is therefore shown to be effective for raising the ranks of the images to which the CNN/SVM system unduly gave low ranks. This result suggests integrating the CNN/SVM and fRCNN/SVM systems, because fRCNN makes it possible to evaluate shots complementary to those ranked higher by the CNN/SVM system.

The CNN/SVM+fRCNN/SVM system could not yield advantages over the conventional system for the concepts that represent the detailed attributes of a human, such as "male person," "female person," "old people," and "infants," and those representing human actions. Since fRCNN-based feature extraction detects these detailed human categories as a broader concept "person," the features obtained were not sufficiently discriminative. In addition, "swimming," "singing," and "walking" are the examples of the concepts representing human actions. For verifying these concepts accurately, it is not sufficient to detect only "person," but it is also required to extract features representing the kinds of human action. Figure 4 shows the keyframe image of the shot that the fRCNN/SVM system ranked 18-th for the

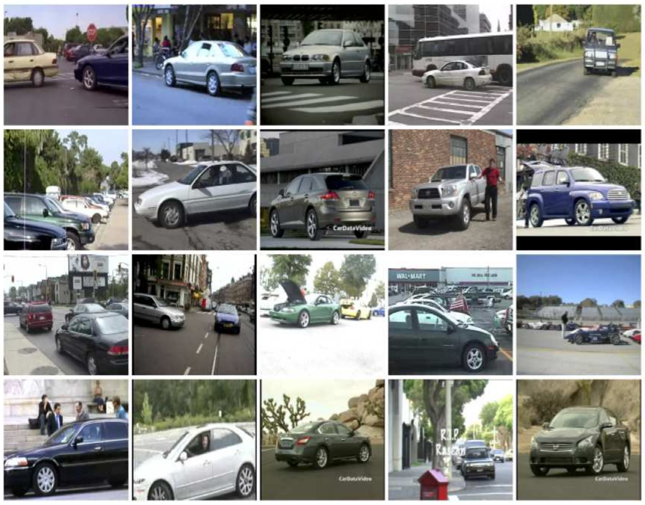

Figure 2. Keyframe images for top 20 shots that CNN/SVM system estimated as "Ground_Vehicles."

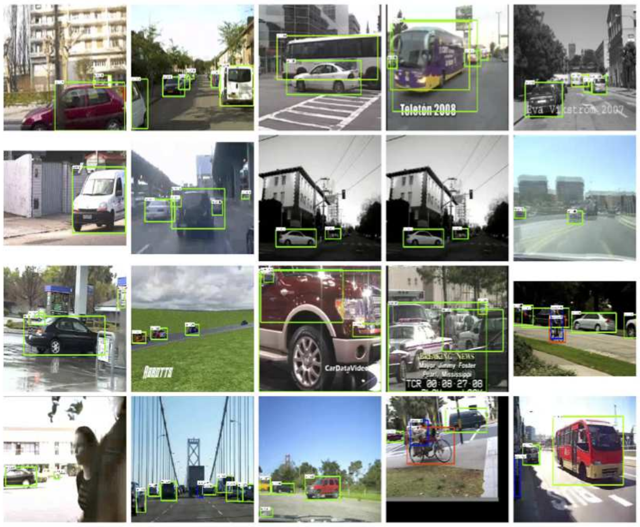

Figure 3. Keyframe images for top 20 shots that fRCNN/SVM system estimated as "Ground_Vehicles."

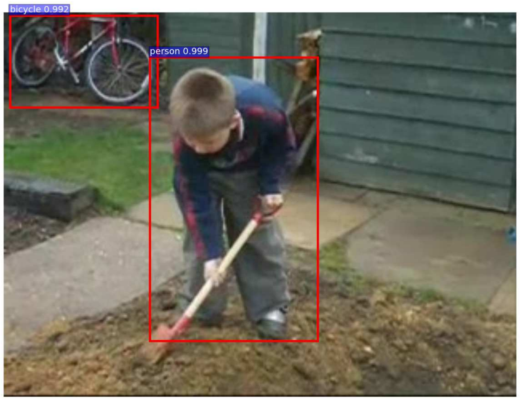

Figure 4. Keyframe image of shot that fRCNN/SVM system ranked 18-th for "Bicycling."

concept "Bicycling." The concept "Bicycling" is defined in TRECVID as "a person riding a bicycle." Although both a bicycle and a person exist in Fig. 4, this image does not fit "Bicycling" because the person is not riding the bicycle. This indicates that not only the detected objects but also their positional relationship are important for some concepts.

4. Conclusion

Object detection using fRCNN was incorporated into semantic indexing to reduce the harmful effect derived from redundant background information. The proposed fRCNN-derived features represent a combination of meaningful objects detected in a video frame image. The developed system, which fuses the CNN/SVM and fRCNN/SVM tandem systems, yielded improvements as compared to the conventional system for most concepts in the TRECVID 2010 semantic indexing task. In future work, we will explore more suitable combinations of target objects in fRCNN and the representation of the positional relationship of these objects to extract more discriminative features for semantic indexing.

References

[1] N. Inoue, Y. Kamishima, K. Mori, and K. Shinoda, "TokyoTechCanon at TRECVID 2012," *Proc. TRECVID*, 2012.

[2] C. G. M. Snoek, S. Cappallo, D. Fontijne, D. Julian, D. C. Koelma, P. Mettes, K. E. A van de Sande, A. Sarah, H. Stokman, and R. B. Towal, "Qualcomm Research and University of Amsterdam at TRECVID 2015: Recognizing concepts, objects, and events in video," *Proc. TRECVID*, 2015.

[3] K. Ueki, and T. Kobayashi, "Waseda at TRECVID 2015: Semantic indexing," *Proc. TRECVID*, 2015.

[4] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR)*, pp. 321-330, 2006.

[5] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quéenot, and R. Ordelman, "TRECVID 2015 – An overview of the goals, tasks, data, evaluation mechanisms and metrics," *Proc. TRECVID*, 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Neural Information Processing Systems (NIPS)*, pp. 91-99, 2015.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3 , pp. 211-252, 2015.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717-1724, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Neural Information Processing Systems (NIPS)*, pp. 1097-1105, 2012.

[10] 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.

[11] P. Over, G. Awad, M. Michel, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. Quéenot, "TRECVID 2010 - An overview of the goals, tasks, data, evaluation mechanisms, and metrics," *Proc. TRECVID*, 2010.

[12] S. Ayache, and G. Quénot, "Video corpus annotation using active learning," *Advances in Information Retrieval*, pp. 187-198, 2008.

[13] J. Blanc-Talon, W. Philips, D. Popescu, P. Scheunders, and P. Zemcik, "Advanced concepts for intelligent vision systems," 2012.

[14] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] https://github.com/ShaoqingRen/faster_rcnn