T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Semantic Indexing for Large-Scale Video Retrieval
Authors	Nakamasa Inoue, Koichi Shinoda
Citation	ITE Transactions on Media Technology and Applications, vol. 4, no. 3, pp. 209-217
Pub. date	2016, 7

Powered by T2R2 (Science Tokyo Research Repository)

 Copyright © 2016 by ITE Transactions on Media Technology and Applications (MTA)

Invited Paper

Semantic Indexing for Large-Scale Video Retrieval

Nakamasa Inoue †, Koichi Shinoda †

Abstract Video semantic indexing, which aims to detect objects, actions and scenes from video data, is one of important research topics in multimedia information processing. In the Text Retrieval Conference Video Retrieval Evaluation (TRECVID) workshop, many fundamental techniques for video processing have been developed and have been shown to be effective for real data such as Internet videos. They include extensions of deep learning techniques and image recognition techniques such as bag of visual words to video data. This paper reviews TRECVID activities with these techniques for semantic indexing. We also show the TokyoTech system using Gaussian-mixture-model (GMM) supervectors and deep convolutional neural networks (CNNs) with its experimental evaluation at TRECVID 2014.

Key words: Video Semantic Indexing, Semantic Concept Detection, Video Retrieval, Deep Learning, Bag of Words, Gaussian Mixture Models, TRECVID Workshop

1. Introduction

To communicate with each other from all over the world, video has become a popular choice on the Internet. For example, in many online services such as YouTube, users can easily upload videos to show their activities. Applications using web cameras are developed for online meeting and surveillance. As a result, a huge amount of video data has been made available on Internet archives.

The role of video search engines is to provide relevant videos based on a query given by a user. For example, a person who wants to practice dance will be satisfied if a video segment about choreography for the dance is provided by search engines. To find such relevant video segments, metadata such as tags describing the contents should be attached to each video segment.

Most of recent search engines use text information such as titles and short descriptions of video attached by users. However, this information is often not enough to search specific objects, locations, and scenes, because detailed tags are missing. Since attaching detailed text tags manually is costly, automatic detection of semantic concepts including objects, events and scnes is demanded for video retrieval.

To promote development of video retrieval techniques, the U.S. National Institute of Standards and Technology (NIST) sponsors Text Retrieval Conference Video Retrieval Evaluation (TRECVID) workshop[1]. In TRECVID, many research teams from academic and industries can share video data, video-retrieval tasks, and results.

Semantic indexing has been an important task in TRECVID. It aims to detect objects such as Airplane, actions such as Singing, and scenes such as Cityscape. This task is a challenging task due to the semantic gap, i.e., the lack of relation between low-level video features and high-level semantic concepts[2].

To bridge the semantic gap, most of approaches focus on statistical pattern recognition, where they construct a model for detecting semantic concepts by using labeled training examples. It has been shown that, as the number of training samples increases, more complex models with many parameters improve the detection performance. For example, recent trends are to use probabilistic models such as Gaussian mixture models and neural networks. Their effectiveness has been reported in recent TRECVID workshops with more than one thousand hours of video data shared with participants.

In this paper, we review TRECVID video semantic indexing activities. This paper is organized as follows. Section 2 overviews the TRECVID workshop. Section 3 and 4 show datasets and approaches for the semantic indexing task, respectively. Section 5 shows details of the TokyoTech semantic-indexing system at TRECVID 2014 with its experimental results. Section 6 describes

Received November 25, 2015; Accepted December 15, 2015

† Tokyo Institute of Technology
(2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552, Japan)

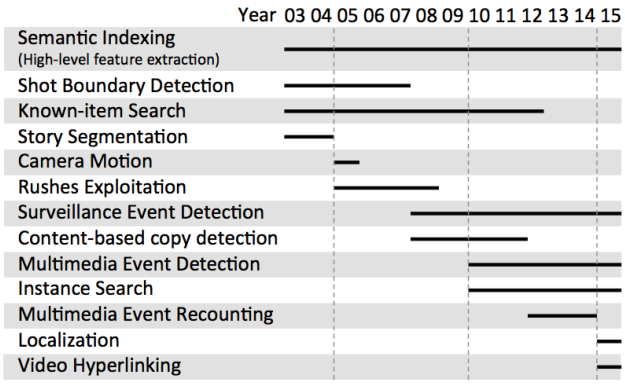

Fig. 1 Overview of the TRECVID tasks from 2003 to 2015. Black bars show years by tasks.

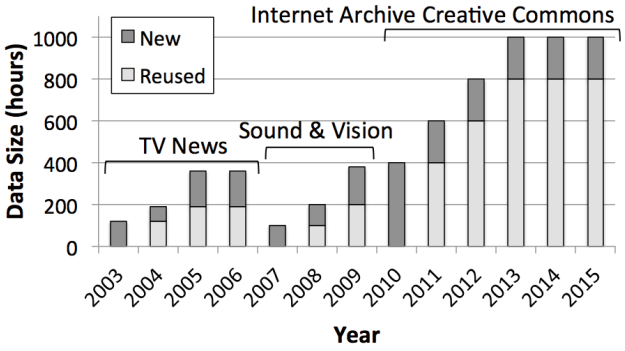

Fig. 2 Size of data used in the Semantic Indexing task.

challenges in the future. Finally, Section 7 concludes this paper.

2. TRECVID Workshop

TRECVID[1] is an annual international workshop sponsored by the National Institute of Standards and Technology (NIST). It began as a track of Text REtrieval Conference (TREC) in 2001, and became an independent workshop from 2003.

As shown in Figure 1, four to six tasks about video retrieval are opened to participants every year. Most of them are challenging tasks using real data such as broadcast, Internet, and surveillance video. Since large-scale video datasets and evaluation measures are shared, it is easy to compare results obtained by different systems.

Compared with image recognition challenges such as ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[59] and PASCAL Visual Object Classes Challenge[60], TRECVID is more focusing on tasks for real applications. Since TRECVID tasks have scenarios for video search, editing, summarizing, surveillance etc., developing automatic and/or interactive systems is required in addition to developing video features and classifiers.

The number of participating teams for each year is fifty to seventy from academic and industries. Each team is required to submit its results at least for one task to attend the workshop.

Recent long-run tasks are Semantic Indexing (SIN), Multimedia Event Detection (MED), Instance Search (INS), and Surveillance Event Detection (SED).

The SIN task aims to detect semantic concepts including objects, actions, and scenes. Examples of the semantic concepts are "Airplane", "Cityscape", and "Dancing". Many techniques have been developed from this task and they are applicable to the other tasks.

The MED task focuses on detecting complex events described by a short sentence such as "batting in a run". Since an event is often described by a combination of semantic concepts, outputs of a SIN system, such as detection scores for each semantic concept, are helpful for this task[3].

The INS task is for delimiting a specific person, object or place. A query is by an image in this task. For example, given an image of a car, the goal is to find exactly the same car from videos in database. Visual features developed in the SIN task are often applied to this task with feature matching techniques such as BM25[5] and asymmetrical dissimilarity[4]

The SED task aims to detect actions by surveillance cameras at a specific place such as an airport. Compared with the other tasks, detailed analysis such as multiple person tracking is often needed since the main purpose of this task is for improving security[6][7].

In the following, we review datasets and approaches for the SIN task. Notably, many techniques developed in the SIN task can be applied to the other tasks since various semantic concepts are targeted in this task.

3. Datasets and Measures

3.1 Video data

Three types of video datasets are used in the SIN task: Television News (TV), Sound and Vision (SV), and Internet Archive Creative Commons (IACC). As summarized in Figure 2, several hundred hours of video clips are added every year.

Television News (TV)

The TV dataset[8] used from 2003 to 2006 consists of 400 hours of news videos. It includes news in three languages: English (*ABC World, CNN Headline, C-SPAN, NBC*), Arabic (*LBC*), and Chinese (*CCTV4, NTDTV*). Since news videos are edited by professional

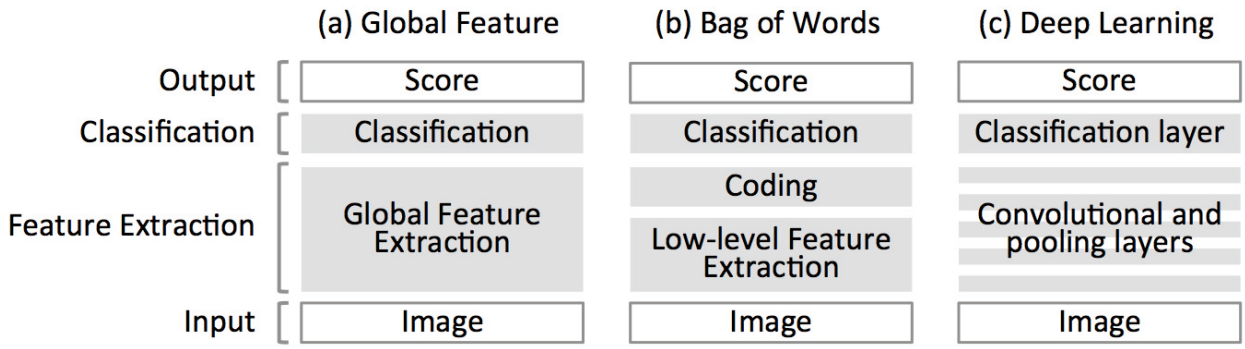

Fig. 3 Three approaches for semantic indexing.

editors, the main object is typically in the center of an image, and its short text description is often shown at the bottom or top.

Sound and Vision (SV)

The SV dataset used from 2007 to 2009 consists of 400 hours of news, documentaries, and educational programming provided by the *Netherlands Institute for Sound and Vision*. Compared with the TV dataset, this dataset has a large diversity. For example, differences in illumination, background, and camera distance/angles often make it difficult to detect semantic concepts.

Internet Archive Creative Commons (IACC)

The IACC dataset, which consists of 1,000 hours of Internet video, is one of the largest datasets for video recognition. This dataset is used from 2010. Video clips are collected from Internet archives with creative commons licenses. Compared with the SV dataset, the quality of video is often low since most of video clips are generated by consumers.

3.2 Annotations

Shot boundary information[9] is provided for each video clip. It splits a video clip into 5-to-10 seconds video shots.

For each video shot, positive or negative labels are attached for each semantic concept. Here, a semantic concept is an object such as "Airplane", an action such as "Dancing", or a scene such as "Cityscape".

346 types of semantic concepts are selected at TRECVID with its basic selection criteria, which requires semantic concepts to 1) be moderately frequent (positive in 1.0% on average), 2) have clear definition, and 3) be of use in searching. Here, definitions of each semantic concept are given by short text descriptions. For example, the definition of "Cityspace" is "View of a large urban setting, showing skylines and building tops".

From 2007, a collaborative annotation system based on active learning[10] is introduced to generate labels. In 2007, 711,566 labels for training data are produced by 32 TRECVID teams participated in collaborative annotation.

3.3 Evaluation Measures

Overall performance is measured by Mean Average Precision (Mean AP), which is the geometric mean of APs over targeted semantic concepts. AP is a value of the area under the recall-precision curve, which is given by

$$AP = \frac{1}{R}\sum_{r=1}^{N} \mathrm{Pre}(r)\mathrm{Rel}(r), \tag{1}$$

where R is the number of positive samples, N is the number of testing shots, $\mathrm{Pre}(r)$ and $\mathrm{Rel}(r)$ are precision and the label (1 for positive, 0 for negative) at the rank r. In TRECVID evaluation, Inferred Average Precision (InfAP)[11], which estimates AP from a subset of labels on testing data, is introduced to reduce the cost of annotation.

4. Semantic Indexing

Most of approaches to semantic indexing are based on statistical pattern recognition. From an input video segment, a semantic-indexing system computes detection confidence scores for each semantic concept.

In TRECVID Semantic Indexing from 2004 to 2014, three main approaches 1) global features, 2) bag of words, and 3) deep learning have been shown to be effective. Notably, the main framework consists of feature extraction and classification as shown in Figure 3 is common among them. In this section, we review studies related to these approaches and fusion of multiple systems.

4.1 Global Features

To detect semantic concepts visible in a video such as "Airplane" and "City", visual features including color and texture information are effective. The idea of global

features is to represent an image by a statistical feature vector such as a histogram. For example, color histogram[12] and color moments[13] extract features from each channel of a color space such as RGB and HVS[14]. Gabor filter banks[15] and edge direction histogram[16] extract textures and edges. These features are invariant to simple transformations such as shifting, and linear scaling. However, they are not robust against illumination changes and viewing angle changes.

With these features, supervised learning techniques are introduced to detect semantic concepts from each video shot. Almost all of recent works focus on supervised learning techniques to construct a detector based on training samples. Support Vector Machine (SVM)[17] is the most commonly used technique to train a discriminative model. Logistic regression[18] and probabilistic output of SVM scores[19] are known to be effective to obtain probabilistic scores. See surveys by Snoek[20], Jain[21], Xu[22] for details of video retrieval and supervised learning techniques.

4.2 Bag of Words

From 2004, bag-of-words methods[23] are used with heuristic low-level features. Its basic idea is to represent an image by an aggregation of local descriptors. For example, in a car image, local descriptors extracted from wheels, headlights, doors, and body line of the car can be useful to represent the car.

A bag-of-words method often consists of two steps: 1) low-level feature extraction and 2) coding to obtain aggregated representation. In the first step, a set of gradient or color descriptors are extracted from an image. Scale Invariant Feature Transform (SIFT)[24] is the most widely used features, which extracts histograms of gradients from each interest point. Histogram of Oriented Gradients (HOG)[25][26] for tracking a person, which can be viewed as simplified SIFT, is also introduced instead of SIFT to reduce computational costs. Color SIFT[27] extends SIFT to color spaces such as RGB and Opponent spaces. These low-level features are extracted by using key-point detectors. To detect objects such as cars, Harris-Laplace detector[28] is applied to extract corner points. To detect actions, space-time detectors such as Space-Time Interest Points[29] and Dense trajectories[30] are also introduced in addition to it. To detect scenes, dense sampling[31][32] for extracting features from grid points has been shown to be effective.

The second step is coding of low-level features. The simplest method is histogram coding[23], which is the zero order statistics to count the number of low-level features for each pre-defined bins designed by applying vector quantization[40]. Soft assignment techniques such as sparse coding[35][36][37], Kernel codebook[38], and Gaussian mixture models[39] are often introduced to reduce the quantization errors. Super-vector coding[31] and VLAD[41] are their extension to use the first order statistics. Fisher-vector[34][33][32] and GMM supervecotr[42][43][44] use the first and second order statistics by ingroducing GMMs.

To model spatial characteristics, spatial pyramid pooling[45][46], which splits an image into 2x2 and/or 3x1 regions, is applied to the above coding techniques. This is effective to classify an object and its background. For example, in an image of a car, a road and sky are often at the bottom and the top of the image, respectively.

To improve the speed of coding, tree-structures such as kd-trees[47][48], metric trees[49][51][50], approximate nearest neighbors[52][53], and random forests[54][55][56] are introduced to the vector quantization step.

4.3 Deep Learning

From 2013, deep learning techniques are introduced to train low-level features automatically from a large set of training data. Especially, recent works focus on deep convolutional neural networks (CNNs). For example, Alex Net[57][58], a network with seven hidden layers, have been shown to be effective for image recognition at the ImageNET[60][59] Large Scale Visal Recognition Challenge (ILSVRC) in 2012. This is an extension of traditional neural networks[61] to a large-scale network with 60 million parameters. To train the parameters, 1.2 million images for 1,000 objects on the ImageNET dataset are used. Some recent works are focusing on lager and deeper networks such as Very Deep Convolutional Networks[63] with 19 laygers, GoogLeNet[64] with 22 layers, and combination with Fisher kernels[65][66][67][68][69] These networks outperforms the AlexNet on the ImageNET dataset.

To semantic indexing on TRECVID, networks pretrained on the ImageNET are often introduced. The simplest way to introduce a pre-trained network is to use activation values at the last or second last fully connected layer as a feature vector, which can be an input vector of a support vector machine (SVM). Note that the network parameters are trained on ImageNET and SVM parameters are trained on TRECVID to save computational costs in this case. Fine-tuning techniques to update all network parameters are effective to further improve the performance. Compared with the bag-of-word methods, deep learning is often more effective for

detecting objects, and less effective for detecting actions[70]. For actions, motion features such as dense trajectories[30] are known to be effective.

4.4 Audio Features

To detect concepts related to audio such as "Singing", "Speaking", and "Dancing", audio features are shown to be effective. For example, Mel Frequency Cepstral Coefficients (MFCCs) are introduced to capture audio information[72][71]. The Fisher vector framework[32] is often applied to obtain audio representations by replacing inputimage descriptors with audio MFCCs. To capture information from speech, automatic speech recognition (ASR) systems are applied to video data. It is known that ASR outputs are effective to detect concepts from news videos[8], in which an announcer speaks about the important topics in video clearly.

4.5 Fusion of Multiple Systems

To capture various semantic concepts, fusion of different systems often improve the detection performance. For example, the best system at the TRECVID 2014 is a hybrid system[70] of bag-of-words and deep learning.

Late fusion, which combines scores obtained from multiple systems, is the easiest way to combine different systems. For example, a weighted sum of SVM scores obtained by bag-of-words and deeply learned features is used as the final score for detection[70]. Weight coefficients are often optimized by using cross-validation.

Early fusion combines features or kernels before training models. For example, by concatenating histogram of oriented gradients (HOG) and histogram of optical flow (HOF) features at low-level feature extraction step in bag-of-words, shapes and movements are captured simultaneously[30]. Multiple kernel learning[73][74][75][76] provides a way to combine systems by taking weighted sum of kernel or distance matrixes. For example, χ^2 kernel[77][78], RBF kernel, histogram intersection kernel [79][80], Fisher kernel[34] have been shown to be effective with bag-of-words representations. These kernels can be inputs to MKL for training weight coefficients with model parameters for detection. Since kernel methods are computationally expensive, linear kernels or linear homogeneous kernel maps[81][82] are often introduced to improve the speed of the testing phase.

5. TokyoTech Semantic Indexing System

Here, we describe the TokyoTech system[71] in TRECVID 2014 Semantic Indexing. It is a hybrid system of GMM supervectors and Deep CNN as shown in Figure 4.

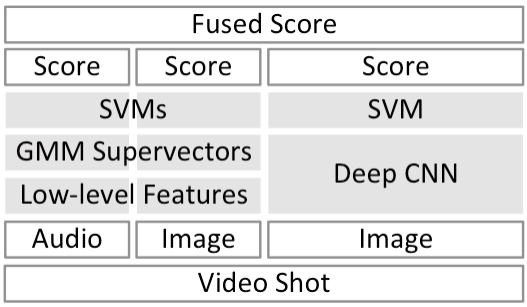

Fig. 4 Overview of our semantic indexing system.

5.1 GMM Supervector

The GMM supervector is an extension of bag-of-visual-words to a probabilistic model. Its idea is to measure distance between two sets of image or audio descriptors by Kullback Leibler divergence. Here, we extract the following six types of descriptors: 1) Harris-Affine SIFT, 2) Hessian-Affine SIFT, 3) Dense SIFT, 4) Dense HOG, 5) Dense LBP[26], and 6) Audio MFCCs.

Let X be a set of descriptors extracted from a video shot. The probability density function of a GMM is given by

$$p(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2}$$

where $x \in X$ is a descriptor, K is the number of mixtures, $\mathcal{N}$ is the probability distribution function of Gaussian distribution, w_k, μ_k, and Σ_k are the k-the weight coefficient, mean vector, and covariance matrix, respectively. The maximum a posterior (MAP) adaptation technique is used to estimate the parameters of a GMM since the number of descriptors in a video shot may not be enough to estimate parameters precisely. With MAP adaptation, a GMM called the universal background model (UBM) is estimated from all training videos first, and then only mean vectors are updated for each shot by assuming weight coefficients and covariance matrixes are common among all video shots. This assumption makes it easy to compute Kullback Leibler divergence between two Gaussian distributions.

From estimated two GMMs p and p', distance between them are defined by the weighted sum of KLD between corresponding Gaussian distibutions as

$$d(p, p') = \sum_{k=1}^{K} w_k \mathrm{KLD}(p_k||p'_k) \tag{3}$$

$$= \sum_{k=1}^{K} w_k (\mu_k - \mu'_k)^T \Sigma_k^{-1} (\mu_k - \mu'_k) \tag{4}$$

where p_k and p'_k are the k-the Gaussian distribution.

A GMM supervector is defined by

$$\phi(X) = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \vdots \\ \tilde{\mu}_K \end{pmatrix}, \quad \tilde{\mu}_k = \sqrt{w_k^{(U)}}(\Sigma_k^{(U)})^{-\frac{1}{2}}\hat{\mu}_k. \quad (5)$$

where $\hat{\mu}_k$ is the estimated mean vector with the MAP adaptation, and $w_k^{(U)}, \Sigma_k^{(U)}$ are the weight coefficient and the covariance matrix for UBM. This GMM supervector is defined so that we obtain

$$d(p, p') = \|\phi(X) - \phi(X')\|_2^2, \quad (6)$$

where X and X' are sets of feature vectors, p and p' are two GMMs estimated from X and X' respectively, and d is the distance given by Eq. (4). This GMM supervector is used as an input of support vector machine to obtain a detection score.

5.2 Deep CNN

The AlexNet[57)58)] with seven hidden layers is introduced to extract features. The parameters of the CNN is trained on ImageNET dataset, which has 1.2 million images for 1,000 object categories.

To detect semantic concepts, support vector machines (SVMs) are trained with labels provided by TRECVID with activation values at the second last layer as its input. The dimension of input vector is 4096. Note that the final layer is not used since it is often over-fitted to training data.

5.3 Late Fusion

Finally, scores obtained by GMM supervectors and Deep CNNs are linearly combined as

$$s = \sum_{f \in \mathcal{F}} \alpha_f s_f, \quad 0 \leqq \alpha_f \leqq 1, \quad \sum_f \alpha_f = 1. \quad (7)$$

where $\mathcal{F}$ is a set of feature types. Combination coefficients α_f are optimized on a validation set.

5.4 Evaluation on TRECVID 2014

In the official evaluation at TRECVID 2014, 660,311 and 107,806 video shots from the IACC dataset are used for training and testing, respectively. Evaluation measure is Mean Inferred Average Precision (InfAP)[11)] among 30 semantic concepts.

Figure 5 compares semantic indexing systems in TRECVID 2014. The system described in Sec. 5 achieves 0.288 Mean InfAP, which was ranked third among participating teams. In 2014, hybrid systems deep learning and bag-of-words performed well. The tendency is as follows. Convolutional neural networks improve the performance for detecting objects such as Airplane, and Car. Bag-of-words with densely samples image descriptors helps to detect scenes such as Cityscape and Forests. Examples of detected video shots are shown in Figure 6. Notably precision at top 10 detected video shots was 80% on average. However, it is still difficult to detect small objects such as basketballs.

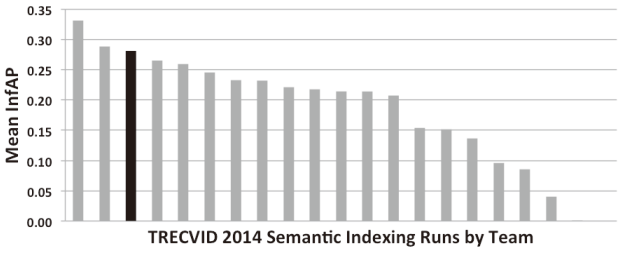

Fig. 5 Comparison of Mean InfAP for Semantic Indexing at TRECVID 2014.

6. Challenges

6.1 More Precisely: Localizing Concepts

Where and when do semantic concepts appear in video? Localizing semantic concepts, aiming to detect their bounding boxes or segmentation masks, is an extended challenging task for semantic indexing.

For object localization, image segmentation methods can be directly applied to video data. For example, selective search[83)84)] is one of the most effective method for detecting bounding boxes. It first uses hierarchy segmentation to detect candidate bounding boxes and then applies classifiers for each candidate. Recent works such as Regions with CNN (R-CNN)[85)] and spatial pyramid pooling for CNN[86)87)] apply selective search to neural-network based frameworks. Scene segmentation for street view images[88)89)90)] could also be effective to segment sky, road, buildings, and objects. Some recent works[91)71)] focus on extending these methods to capture motions in video, for example, by introducing optical flow features for segmentation. However, these methods are not always effective for all types of semantic concepts. For example, actions can not be well localized with them since boundaries of actions are often different from that of objects. For action localization, clear definition of generic action classes to determine boundaries is first needed to be designed. This could help to make large-scale training data for actions.

6.2 More Generally: Expanding Vocabulary

What should we do to make detectors for many types of semantic concepts? The simplest way is to increase the amount of training data. In TRECVID 2010, 500 semantic concepts are defined to be annotated in the collaborative annotation. However, we faced the fact that 150 of them have less than 3 positive samples in

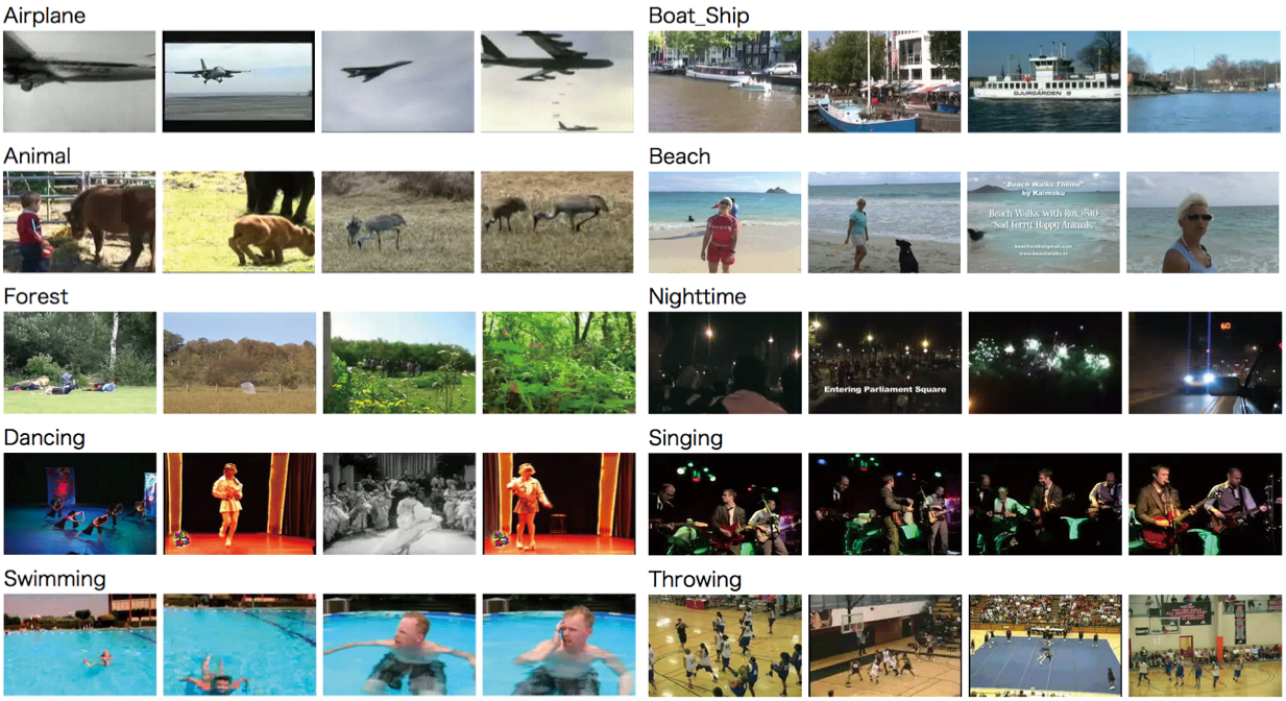

Fig. 6 Top 5 detected video shots for 10 semantic concepts.

200 hours of video data. This shows the difficulty of increasing training samples.

Recently, zero-shot learning has been focused to make detectors for unseen objects or scenes. For example, intermediate presentations such as attributes[92)93)] are introduced to represent an object by them. Word embedding methods[94)95)96)] are introduced to learn relations between semantic concepts on text data and apply them to image or video data, for vocabulary expansion[98)] and sentence generation for images[97)99)]. Textual information such as video title or captions on images is also utilized for zero-shot event detection[100)101)]. In the future, learning methods for integrating different types of multimedia data such as videos, images, audio, and text are needed to make it possible to detect many types of semantic concepts.

7. Conclusion

TRECVID activities and approaches for video semantic indexing are reviewed in this paper. Experimental evaluation on TRECVID 2014 showed the effectiveness of our semantic indexing system using deep learning and GMM supervectors. We conclude that new topics such as localizing concepts video and expanding vocabulary for many semantic concepts will be important challenges in the near future to develop more advanced semantic indexing systems.

References

1) F. A. Smeaton, P. Over and W. Kraaij, Evaluation campaigns and TRECVid. *Proc. ACM MIR workshop*, pp. 321–330, 2006.
2) A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. R. Jain, Content-based Image Retrieval at the End of the Early Years, *In IEEE Trans. on PAMI*, vol. 22, no. 12, pp. 1349–1380, 2000.
3) A. Habibian, T. Mensink, and C.G.M. Snoek. VideoStory: a new multimedia embedding for few-example recognition and translation of events. *Proc. ACM Multimedia*, 2014.
4) C. Z. Zhu, H. Jegou, S. Satoh, Query-Adaptive Asymmetrical Dissimilarities for Visual Object Retrieval, *Proc. ICCV*, pp. 1705–1712, 2013.
5) M. Murata, H. Nagano, R. Mukai, K. Kashino, S. Satoh, BM25 with Exponential IDF for Instance Search *In IEEE Trans. on Multimedia*, vol. 16, no. 6, pp. 1690–1699, 2014.
6) C. Gao, D. Meng, W. Tong, Y. Yang, Y. Cai, H. Shen, G. Liu, S. Xu, and A. G. Hauptmann. Interactive surveillance event detection through mid-level discriminative representation. *Proc. ICMR*, 2014.
7) Q. Chen, Y. Cai, L. Brown, A. Datta, Q. Fan, R. Feris, S. Yan, A. Hauptmann, and S. Pankanti, Spatio-Temporal Fisher Vector Coding for Surveillance Event Detection, *Proc. ACM Multimedia*, 2013.
8) J. L. Gauvain, L. Lamel, and G. Adda, The LIMSI Broadcast News Transcription System. *In Speech Communication*, vol. 37, no. 1, pp. 89-108, 2002.
9) C. Petersohn, Fraunhofer HHI at TRECVID 2004 - Shot Boundary Detection System, *Proc. TRECVID workshop*, 2004.
10) S. Ayache, and G. Quenot, Video Corpus Annotation using Active Learning, *Proc. ECIR*, 2008.
11) Estimating Average Precision with Incomplete and Imperfect Judgments E. Yilmaz, J. A. Aslam, *Proc. CIKM*, 2006.
12) C. L. Novak, and S. A. Shafer, Anatomy of a color histogram, *Proc. CVPR*, pp. 599-605, 1992.
13) M. Stricker, M. Stricker, M. Orengo, and M. Orengo, Similarity of Color Images, *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pp. 381-392, 1995.
14) S. Sural, G. Qian, S. Pramanik, Segmentation and Hitogram Generation Using the HSV Color Space for Image Retrieval, *Proc. ICIP*, pp. 589-592, 2002.
15) B. S. Manjunath, W.-Y. Ma, Texture Features for Browsing and Retrieval of Image Data, *IEEE Trans. on PAMI*, vol. 18, no. 8, pp. 837-842, 1996.
16) A. K. Jain and A. Vailaya, Image Retrieval Using Color and

Shape, *Elsevier Pattern Recognition*, vol. 29, no. 8, pp. 1233-1244, 1996.
17) C. Cortes, and V. Vapnik, Support-Vector Networks, *Springer Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
18) D. R. Cox, The Regression Analysis of Binary Sequences , *In JRSS*, vol. 20, no. 2, pp. 215-242, 1958.
19) J. C. Platt, Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, *Advances in Large Margin Classifiers*, pp. 61-74, 1999.
20) C.G.M. Snoek, and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2009.
21) A. K. Jain, R. P. W. Duin, and J. Mao, Statistical pattern recognition: A review, *IEEE Trans. on PAMI*, vol. 22, pp. 4-37, 2000.
22) R. Xu, and D. Wunsch II. Survey of clustering algorithms. *IEEE Trans. on Neural Networks*, vol. 16(3), pp. 645–678, 2005.
23) G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual Categorization with Bags of Keypoints. *Proc. ECCV SLCV workshop*, pp. 59–74, 2004.
24) D. G. Lowe. Distinctive image features from scale-invariant keypoints. *In IJCV*, vol. 60(2), pp. 91–110, 2004.
25) N. Dalal and W. Triggs. Histograms of oriented gradients for human detection. *Proc. CVPR*, pp. 886–893, 2004.
26) X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. *Proc. ICCV*, pp. 32–39, 2009.
27) K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. on PAMI*, vol. 32(9), pp. 1582–1596, 2010.
28) K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool, A Comparison of Affine Region Detectors, *IJCV*, vol. 65, no. 1, pp. 43-72, 2005.
29) H. Wang, M. M. Ullah, A. Klaser, I. Laptev and C. Schmid, Valuation of Local Spatio-temporal Features for Action Recognition, *Proc. BMVC*, 2009.
30) H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *In IJCV*, vol. 103, no. 1, pp. 60–79, 2013.
31) X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. *Proc. ECCV*, pp. 141–154, 2010.
32) F. Perronnin, S. Jorge, and T. Mensink. Improving the fisher kernel for large-scale image classification. *Proc. ECCV*, pp. 143–156, 2010.
33) F. Perronnin and et al. Fisher kernels on visual vocabularies for image categorization. *Proc. CVPR*, 2007.
34) T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Proc. NIPS*, pp. 487–493, 1998.
35) J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, Locality-constrained linear coding for image classification, *Proc. CVPR*, pp. 3360–3367, 2010.
36) H. Bristow, A. Eriksson, and S. Lucey. Fast convolutional sparse coding. *Proc. CVPR*, pp. 391–398, 2013.
37) T. Guha and R. Ward. Learning sparse representations for human action recognition. *IEEE Trans. on PAMI*, vol. 34(8), pp. 1576–88, 2012.
38) J. C. V. Gemert, J.-m. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. *Proc. ECCV*, pp. 696–709, 2008.
39) F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. *Proc. ECCV*, pp. 464–475, 2006.
40) A. Gersho, and R. M. Gray. Vector quantization and signal compression. *Kluwer Academic Publishers*, 1992.
41) Aggregating Local Descriptors Into a Compact Image Representation, H. Jegou, M. Douze, C. Schmid, and P. Perez, *Proc. CVPR*, pp. 3304–3311, 2010.
42) W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.
43) N. Inoue and K. Shinoda. A fast map adaptation technique for GMM-supervector-based video semantic indexing systems, *Proc. ACM Multimedia*, pp. 1357–1360, 2011.
44) N. Inoue, Y. Kamishima, T. Wada, K. Shinoda, and S. Sato. Semantic indexing using gmm supervectors and tree-structured GMMs (TokyoTech+Canon at TRECVID 2011). *Proc. TRECVID workshop*, 2011.
45) S. Lazebnik, C. Schmid, J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *Proc. CVPR*, pp. 2169-2178, 2006.
46) T. Huang. Linear spatial pyramid matching using sparse coding for image classification. *Proc. CVPR*, pp. 1794–1801, 2009.
47) C. Silpa-anan and R. Hartley. Optimised kd-trees for fast image descriptor matching. *Proc. CVPR*, pp. 1–8, 2008.
48) R. F. Sproull. Refinements to nearest-neighbor searching in k-dimensional trees. *In Algorithmica*, vol. 6(1), pp. 579–589, 1991.
49) P. Ciaccia, M. Patella, and P. Zezula. An Efficient Access Method for Similarity Search in Metric Spaces. *Proc. VLDB*, 1997.
50) S. M. Omohundro. Five balltree construction algorithms. *ICSI Technical Report*, TR-89-063, 1989.
51) J. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *In Elsevier Information Processing Letters*, vol.40(4), pp.175–179, 1991.
52) J. S. Beis and D. G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. *Proc. CVPR*, pp. 1000–1006, 1997.
53) M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *Proc. VISAPP*, pp. 331–340, 2009.
54) F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Trans. on PAMI*, vol. 30(9), pp. 1632–1646, 2008.
55) J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-Time Visual Concept Classification. *IEEE Trans. on Multimedia*, vol. 12(7), pp.665–681, 2010.
56) D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Proc. CVPR*, pp. 2161–2168, 2006.
57) A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Proc. NIPS*, 2010.
58) Y. Jia, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. *Proc. ACM Multimedia Open Source Competition*, 2014.
59) 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-Fe, ImageNet Large Scale Visual Recognition Challenge, *IJCV*, pp. 1–42, 2015.
60) S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman The Pascal visual object classes challenge: a retrospective. *In IJCV*, vol. 111(1), pp. 98–136, 2014.
61) Y. L. Cun, B. Boser, J. S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel, Hand-Written Digit Recognition with a Back-Propagation Network, *Proc. NIPS*, 1990.
62) C.G.M. Snoek, K.E.A. van de Sande, D. Fontijne, A. Habibian, M. Jain, S. Kordumova, Z. Li, M. Mazloom, S.L. Pintea, R. Tao, D.C. Koelma, and A.W.M. Smeulders. The mediamill at trecvid 2013: searching concepts, objects, instances and events in video. *Proc. TRECVID workshop*, 2013.
63) K. Simonyan, and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *Proc. ICLR*, 2015.
64) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich , Going Deeper With Convolutions, *Proc. CVPR*, 2015.
65) K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher betworks for large-scale image classification. *Proc. NIPS*, 2013.
66) V. Sydorov, M. Sakurada, and C.H. Lampert. Deep Fisher kernels – end to end learning of the Fisher kernel GMM parameters. *Proc. CVPR*, 2014.
67) Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. *Proc ECCV*, 2014.
68) J. Y. Ng, F. Yang, and L. S. Davis. Exploiting local features from deep networks for image retrieval. *Proc. CVPR workshop on Deep Vision*, 2015.
69) F. Perronnin, and D. Larlus. Fisher vectors meet neural networks: a hybrid classification architecture. *Proc. CVPR*, 2015.
70) C.G.M. Snoek, K.E.A. van de Sande, D. Fontijne, S. Cappallo, J. van Gemert, A. Habibian, T. Mensink, P. Mettes, R. Tao, D.C. Koelma, and A.W.M. Smeulders. Video concept detection by deep nets with FLAIR (Mediamill at TRECVID 2014: searching concepts, objects, instances and events in video). *Proc. TRECVID workshop*, 2014.
71) N. Inoue, et al., TokyoTech at TRECVID 2015, *Proc. TRECVID workshop*, 2015.
72) Informedia at TRECVID 2014, *Proc. TRECVID workshop*, 2014.
73) G. R. Lanckriet, et al., Learning the Kernel Matrix with Semidefinite Programming, *In JMLR*, vol. 5, pp. 27–72, 2004.
74) F. R. Bach, et al., Multiple Kernel Learning Conic Duality and the SMO Algorithm, *Proc. ICML*, pp. 6, 2004.
75) L. Chen, L. Duan, D. Xu, Event Recognition in Videos by Learn-

ing from Heterogeneous Web Sources, *Proc. CVPR*, pp.2666–2673, 2013.
76) L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank, Domain Transfer Multiple Kernel Learning, *IEEE Trans. on PAMI*, vol. 34, no. 3, pp. 465–479, 2012.
77) J. R. R. Uijlings, A. W. M. Smeulders, R. J. H. Scha, Real-time Bag-of-Words, Approximately, *Proc. CIVR*, 2009
78) K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. *Proc. BMVC*, pp. 1–12, 2011.
79) A. Barla, F. Odone, A. Verri, Histogram Intersection Kernel for Image Classification, *Proc. ICIP*, pp.14–17, 2003.
80) S. Maji, A. C. Berg, and J. Malik, Classification using intersection kernel support vector machines is efficient. *Proc. CVPR*, 2008.
81) A. Vedaldim and A. Zisserman, Efficient Additive Kernels via Explicit Feature Maps, *Proc. CVPR*, 2010.
82) A. Vedaldim and A. Zisserman, Efficient Additive Kernels via Explicit Feature Maps, Efficient Additive Kernels via Explicit Feature Maps. *IEEE Trans. on PAMI*, 2010.
83) K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, A. W. M. Smeulders, Segmentation As Selective Search for Object Recognition, *In Proc. ICCV*, 2011.
84) J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders, Selective Search for Object Recognition, *In IJCV*, vol. 104, no. 2, pp. 154–171, 2013.
85) R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *Proc. CVPR*, 2014.
86) K. He, X. Zhang, S. Ren, and J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, *Proc. ECCV*, 2014.
87) K. He, X. Zhang, S. Ren, and J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, *In IEEE TPAMI*, 2014.
88) J. Xiao, L. Quan, Multiple View Semantic Segmentation for Street View Images, *Proc. ICCV*, pp. 686–693, 2009.
89) J. Yang, B. Price, S. Cohen, and M.-H. Yang, Context Driven Scene Parsing with Attention to Rare Classes, *Proc. CVPR*, pp. 3294–3301, CVPR, 2014.
90) A. Sharma, O. Tuzel, D. W. Jacobs, Deep Hierarchical Parsing for Semantic Segmentation, *Proc. CVPR*, pp. 530–538, 2015.
91) D. Oneata, J. Revaud, J. Verbeek, C. Schmid, Spatio-Temporal Object Detection Proposals, *Proc. ECCV*, pp. 1–16, 2014.
92) D. Jayaraman, K. Grauman, Zero Shot Recognition with Unreliable Attributes, *Proc. NIPS*, 2014.
93) C. H. Lampert, H. Nickisch, and S. Harmeling, Attribute-Based Classification for Zero-Shot Visual Object Categorization, *In IEEE TPAMI*, vol. 36, no. 3, 2014.
94) T. Mikolov, et al. Efficient Estimation of Word Representations in Vector Space. *Proc. ICLR*, 2013.
95) T. Mikolov, et al. Distributed Representations of Words and Phrases and their Compositionality. *Proc. NIPS*, 2013.
96) D. Jacob, et al. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. *Proc. ACL*, 2014.
97) A. Frome, et al. Devise: A deep visual-semantic embedding model. *Proc. NIPS*, 2013.
98) N. Inoue, and K. Shinoda, Vocabulary Expansion Using Word Vectors for Video Semantic Indexing *Proc. ACM Multimedia*, 2015.
99) J. Weston, et al. Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings. *Porc. ECML*, 2010
100) S. Wu,, et al. Zero-shot Event Detection using Multi-modal Fusion of Weakly Supervised Concepts. *Proc. CVPR*, pp.2665–2672, 2014.
101) J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. *Proc. CIKM*, pp.1857–1860, 2013.

Nakamasa Inoue received the B.S., M.S. and D.Eng. degrees in computer science from Tokyo Institute of Technology, Tokyo, Japan, in 2009, 2011, and 2014, respectively. He is currently an Assistant Professor with Tokyo Institute of Technology. His research interests include multimedia information retrieval, statistical pattern recognition, visual and audio categorization, and large-scale benchmark evaluations. He is a member of the IEEE and IEICE.

Koichi Shinoda received the B.S. and M.S. degrees from the University of Tokyo, Tokyo, Japan in 1987 and 1989, respectively, both in physics, and the D. Eng. Degree in computer science from the Tokyo Institute of Technology, Japan, in 2001. In 1989, he joined NEC Corporation, Japan, where he was involved in research on automatic speech recognition. From 1997 to 1998, he was a Visiting Scholar with Bell Labs, Lucent Technologies, Murray Hill, NJ. From June 2001 to September 2001, he was a Principal Researcher with Multimedia Research Laboratories, NEC Corporation. From October 2001 to March 2002, he was an Associate Professor with the University of Tokyo, Japan. He is currently a Professor with the Tokyo Institute of Technology. His research interests include speech recognition, video information retrieval, statistical pattern recognition, and human interfaces. He received the Awaya Prize from the Acoustic Society of Japan in 1997 and the Excellent Paper Award from the IEICE in 1998. He was Publicity Chair in INTERSPEECH2010, Video Program Co-Chair in ACM Multimedia 2012. Dr. Shinoda is a senior member of IEEE, IEICE. He is a member of ACM, IPSJ, JSAI, and ASJ. He is currently an associate editor of Computer Speech and Language and Speech Communication, Elsevier.