Semantic classifier based on compressed sensing for image and video annotation

G. Ding and K. Qin

A new semantic classification approach for image and video annotation is proposed, which fits a semantic classification task into theory of a compressed sensing framework. The proposed approach first utilises training samples to create a dictionary matrix and then uses a matching pursuit algorithm to find the sparse vector. The final annotations are determined according to the reconstruction value from the positive samples and the sparse vector. A systematic performance study on TRECVID 2008 video dataset and Corel image dataset shows the proposed approach is more effective than the traditional support vector machine scheme.

Introduction: With the development of network technology, data compression technology and digital photography technology, the number of images and videos available on the Internet has exploded. It has created a compelling need for innovative tools to retrieve and manage these images and videos. One major challenge is to bridge the so-called 'semantic gap' between low-level features and high-level semantic concepts. Recent studies reveal that semantic annotation for image or video is a promising approach to bridging the gap. Existing automatic image and video annotation methods are mainly based on binary classification models, which try to automatically assign concepts onto image or video by learning the relationships between visual features and concepts. The National Institute of Standards and Technology (NIST) has established 'high-level feature extraction' as a task in TREC video retrieval evaluation (TRECVID) since 2003 [1]. Most of the methods submitted are based on support vector machines (SVMs), which determine the correspondence between concepts and videos as follows. Videos are first segmented into shorter units, such as shots and sub-shots. Low-level features are then extracted from each unit to describe its content. Video annotation is then formalised to learn a set of predefined concepts for each unit based on these low-level features [2]. Currently, the classification accuracy of these methods still cannot meet the requirements of large-scale semantic annotation. Therefore, it is very important to explore new technology to solve the automatic semantic annotation problem. The theory of the compressed sensing (CS) framework has recently attracted research attention because of its superior performance in signal acquisition and representation. The application of CS on image classification has also been studied. In [3], a face recognition algorithm is proposed based on CS and showed better performance compared to the traditional approach. However, this method is not suitable for large-scale semantic annotation. First, the method utilised training samples of all semantic concepts to create a unified dictionary matrix, which will result in substantial increase in computational complexity with the increase in the number of semantic concepts. Secondly, this method did not use any negative samples which is very important for semantic annotation. Based on the above considerations, we propose a new semantic classifier based on CS for image and video annotation, which generates an independent dictionary matrix for each semantic concept according to its training samples. In the training phase, all positive and negative samples are used to determine the classification threshold of this concept. Experimental results demonstrate the proposed approach is promising for semantic annotation tasks.

Compressed sensing: In this Section, we briefly introduce the theory of the CS framework used in the Letter. CS builds upon a core tenet of signal processing and information theory. It employs a nonadaptive linear projection that preserves the structure of the signal; the signal is then reconstructed from these projections using an optimisation programming process. For a 1D discrete signal x, its projection can be expressed as $x = \Psi s$, where x and s are $1 \times \mathrm{N}$ and $\mathrm{K} \times 1$ column vector, and dictionary matrix $\Psi = [\Psi_1, \Psi_2, \Psi_k] \in R^{n \times k}$ $(\mathrm{n} < \mathrm{k})$. The signal x has a sparse representation if it is a linear combination of only M basis vectors. That is, only M coefficients of $\{s_i\}$ $(i = 1, 2, \ldots, \mathrm{N})$ are nonzero and the rest are zero. In CS, signal x is compressed by a projection matrix $\Phi \in R^{M \times N}$ with $(\mathrm{M} < \mathrm{N})$, which yields the sensing vector $y = \Phi x = \Phi\Psi s = \Theta s$. Since $\mathrm{M} < \mathrm{N}$, the task of recovering s from y is underdetermined. However, the additional assumption of the sparsity of s makes it possible and practical. The reconstruction problem can be written as:

$$\hat{s} = \arg\min \|s\|_1, \quad s.t. \, y = \Theta s \tag{1}$$

After obtaining $\hat{s}$, the reconstructed signal is calculated as $\hat{x} = \Psi\hat{s}$. Matching pursuit (MP) [4] is one of the most effective methods for finding sparse representation $\hat{s}$ to reconstruct original signal x.

CS-based semantic classifier (CSSC): Motivated by CS used to effectively implement face recognition [3], we present a semantic classifier based on CS for image and video annotation, referred to as CSSC. In CSSC, for each concept, we use low-level features of its training images to create dictionary matrix $\Psi = \{I_P, I_N\} \in R^{n \times k}$, where low-level features of each image is an n-dimensional vector and k is the total number of training images. I_P and I_N represent the positive samples and negative ones, respectively. Projection matrix Φ is chosen as a random matrix with i.i.d. Gaussian entries. Then Θ can be calculated as $\Theta = \Phi\Psi$. Given an input image or keyframe, the projected sample $y = \Phi x$ is calculated, and then MP is performed to find sparse vector $\hat{s}$. The input sample is final annotated to be 'positive' if the reconstruction error $e = \|x - [I_P, 0] \cdot \hat{s}\|$ is smaller than the predefined threshold T_θ. Otherwise, it is annotated to be 'negative'. The proposed approach is summarised in detail in algorithms 1 and 2.

Algorithm 1: CSSC-training phase

Input: Training dataset: $T^+_{C_i}$ and $T^-_{C_i}$ for concept c_i, where $T^+_{C_i}$ and $T^-_{C_i}$ represent the positive sample set and the negative sample set, respectively.

Output: Threshold for concept c_i : $T^{c_i}_\theta$

(1) Select L positive samples x^+_j and L negative samples x^-_j from $T^+_{C_i}$ and $T^-_{C_i}$ for learning $T^{C_i}_\theta$

(2) Create dictionary matrix $\Psi_{C_i} = [A^+_{C_i}, A^-_{C_i}]$, where $A^+_{C_i}$ and $A^-_{C_i}$ are the rest samples of $T^+_{C_i}$ and $T^-_{C_i}$

(3) Set random projection matrix $\Phi \in R^{M \times N}$ with normalised rows.

(4) For j = 1 to L

a) Calculate $y = \Phi x^+$

b) Use MP to find sparse vector $\hat{s}$ subject to $y = \Phi\Psi_{c_i}\hat{s}$

c) Calculate reconstruction error: $e^+_j = \|x^+_j - [A^+_{c_i}, 0]\,\hat{s}\|$

d) Calculate $y = \Phi x^-_j$

e) Use MP to find sparse vector $\hat{s}$ subject to $y = \Phi\Psi_{c_i}\hat{s}$

f) Calculate reconstruction error: $e^-_j = \|x^-_j - [A^-_{c_i}, 0]\,\hat{s}\|$

(5) Output $T^{c_i}_\theta = 1/2\left((1/L)\sum_{j=1}^{L} e^+_j + (1/L)\sum_{j=1}^{L} e^-_j\right)$

Algorithm 2: CSSC-annotating phase

Input: Given low-level feature of new input image x

Output: x is positive or negative for c_i

(1) Calculate $y = \Phi x$

(2) Use MP to find sparse vector $\hat{s}$ subject to $y = \Phi\Psi_{c_i}\hat{s}$

(3) Calculate reconstruction error: $e = \|x - [A^+_{c_i}, 0]\,\hat{s}\|$

(4) Output x is positive for c_i if $e < T^{c_i}_\theta$, otherwise x is negative for c_i

Experimental results: To evaluate the proposed approach for image and video annotation, we conducted the experiments on two datasets: the benchmark video corpus of the TRECVID 2008 dataset and the benchmark image corpus of Corel dataset. In TRECVID 2008 dataset, there are about 200 hours of broadcast news videos [1]. These training videos are segmented into 35 766 video shots (keyframes) and 20 concepts are labelled on each shot (keyframe). In the experiments, about 100 hours of videos are used as training data. Other videos are used as test data. In Corel dataset, we selected 10 concepts, including beach, building, bus, dinosaur, eagle, flowers and so on, as our images database. Each concept contains 100 images. In the experiments, 50 images of each concept are used as training data; the other images are used as test data. In this Letter, we used Color Histograms (72-dimensional vector) and Wavelet Texture (48-dimensional vector) as the low-level features of the keyframe or image. M of the projection matrix is set to 120, which is equal to the number of low-level feature's dimensions. The value of L is set to 10 in algorithm 1. In the following, we make a comparison between our approach and the SVM-based approach. To implement the SVM-based approach, we use LibSVM [5], which is a simple and efficient software for SVM classification. In LibSVM, we set SVM type to 'C-svc', and the kernel to 'radial basis function'. The optimal parameters of LibSVM can be automatically estimated.

For performance evaluation of video annotation, we use inferred average precision (AP) [6] as the performance metric, which is the official performance metric in TRECVID. It reflects the performance on multiple average precision values along a precision-recall curve. We average the APs over all 20 concepts to create the mean average

precision (MAP), which is the overall evaluation result. Fig. 1 illustrates the AP of the two approaches. We can see that the proposed approach (CSSC) outperforms the SVM-based approach for 11 of all 20 concepts. The MAPs of the two approaches are 0.03772 and 0.0359, respectively. The CSSC has an improvement of 5.04% over the SVM-based approach. For performance evaluation of the image annotation, we use precision as the performance metric. Table 1 shows the annotation results of the images by two approaches. From Table 1, it can be seen that the precisions of the CSSC are higher than that of the SVM-based approach for all concepts. All these comparisons demonstrate the CS-based semantic classifier is another promising solution for image and video annotation tasks.

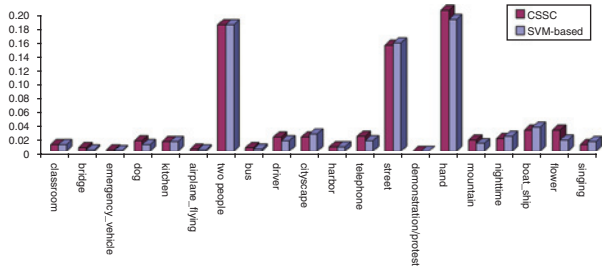

Fig. 1 *Comparisons of AP of two approaches over 20 concepts for video annotation task*

Table 1: Comparisons of precision of two approaches over 10 concepts for image annotation task

	Beach	Building	Bus	Cat	Dinosaur	Eagle	Elephant	Horse	Waterfall	Flower
CSSC	65%	62%	71%	75%	98%	81%	63%	52%	65%	73%
SVM-based	46%	54%	68%	64%	98%	72%	52%	40%	58%	66%

Conclusions: In this Letter, an approach for image and video annotation based on the theory of CS framework is presented. The experiments conducted on the TRECVID dataset and the benchmark image dataset demonstrate that the proposed approach is effective and promising for the semantic image and video annotation tasks.

Acknowledgments: The authors acknowledge the support received from the National Natural Science Foundation of China (project 60972096) and National 863 Plans Projects (grant 2009AA01Z410).

8 August 2009
doi: 10.1049/el.2010.2295
One or more of the Figures in this Letter are available in colour online.

G. Ding and K. Qin (*School of Software, Tsinghua University, Beijing 100084, People's Republic of China*)

E-mail: dinggg@tsinghua.edu.cn

References

1 TREC Video Retrieval Evaluation (TRECVID). http://www-nlpir.nist.gov/projects/trecvid/
2 Tang, J., Hua, X.-S., Mei, T., Qi, G.-J., and Wu, X.: 'Video annotation based on temporally consistent Gaussian random field', *Electron. Lett.*, 2007, **43**, (8), pp. 448–449
3 Vo Nhat, Vo Duc, Challa, S., and Moran, B.: 'Compressed sensing for face recognition'. Presented at IEEE Symp. on Computational Intelligence for Image Processing, Nashville, TN, USA, March–April, 2009, pp. 104–109
4 Mallat, S., and Zhang, Z.: 'Matching pursuit with time-frequency dictionaries', *IEEE Trans. Signal Process.*, 1993, **42**, (12), pp. 3397–3415
5 C.C. Chang, C.J. Lin, 'LIBSVM: a library for support vector machines' (2001). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
6 Aslam, J.A., Pavlu, V., and Yilmaz, E.: 'Statistical method for system evaluation using incomplete judgments'. Proc. 29th ACM SIGIR Conf., Seattle, WA, USA, 2006