Sparse Feature Representation for Visual Tracking

Yifei Liu, Zhenjun Han, Qixiang Ye, Jianbin Jiao, Ce Li
Pattern Recognition and Intelligent System Development Laboratory
Graduate University of Chinese Academy of Sciences
Beijing, China

Abstract—**In this paper, a novel sparse feature representation method for object tracking is proposed. The method is on the observation that a tracked object can be dynamically and compactly represented by a few features (sparse representation) from a large feature set (the improved histogram of oriented gradient and color, HOGC). Based on the HOGC features, the sparse representation can be learned online from the constructed training samples during the tracking procedure by exploiting the L1-norm minimization principle, which can also be called feature selection procedure, ensuring the tracking can adapt to the appearance variations of either foreground or background. Experiments with comparisons demonstrate the effectiveness of the proposed method.**

Keywords*- online feature selection; sparse representation*

I. INTRODUCTION

Visual tracking has been one of the hottest and most significant topics in computer vision for more than two decades, and it is a prerequisite for video data understanding. The aim is to automatically locate the specific object in digital movies after the object's location is initialized.

The previous research on object tracking falls into three different categories: appearance modeling, motion modeling, and searching methods [1]. Motion models are employed to predict the object's location in a new frame based on its history motion characteristics. This can improve the tracking stabilization and make the tracking survive some occlusions [2]. Given a tracked object, searching methods use various matching strategies to find its position in a new video frame. In addition, when the object varies in size, it is necessary to calculate the scale parameter [3].

Motion models and searching methods has been developed by many researchers and have achieved some success, in which filters like Kalman filter[4] and Particle filter[5], Mean-shift method[6] for searching are the most popular methods. Although motion models and searching algorithms are crucial to object tracking, visual tracking is still a challenging topic, especially when the appearances of the tracked object and scene background dynamically change during tracking. This improves the fact that the most important issue in object tracking is whether the object representation is effective enough during all the tracking process, and it has not been solved properly. Many objects are represented by color histogram (HC) features [7] or histogram of oriented gradient (HOG) features [8] for their effectiveness in visual tracking and human detection systems. However, HC cannot work well when the object and its background have the similar color, while HOG, lower efficiency compared with HC, cannot represent objects effectively when the contours of them are indistinctive, especially when objects have large smooth regions. Han, Ye and Jiao [9] integrated the HOG and HC features together (named the histogram of oriented gradient and color, HOGC) for object tracking considering their mutual complementary. However, instead of calculating the global HC and local HOG for the whole object region, we improve the HOGC features by calculating its regional (local) HC and HOG.

Recently, treating tracking as a classification problem, that distinguishes the object from its background based on the object representation during the tracking procedure, has gained much attention [10]. Then the central issue in object tracking becomes which features are important and distinctive for tracking. Investigation in the human vision system (HVS) has shown that a small selective subset of neurons is active for a variety of specific stimuli [11], such as color, texture, scale and contour, etc, and number of active neurons is really sparse in human vision system. Based on the study of HVS, sparse representation, using L1-norm minimization for feature selection that originates from compressed sensing theory [12], has attracted increasing interests in computer vision research community in recent years, and it has been successfully applied in the field of face recognition [13] , object tracking [14] and human detection [15] etc, validating the effectiveness of the representation method. Therefore, we exploit the insight behind the sparse representation, that the features which can most compactly distinguish the object from its background are more important and should be selected for tracking, and online training samples will be constructed to calculate the sparse representation and updated to avoid the drift problem in the tracking procedure.

In this paper, an online training sample set is firstly constructed for sparse feature selection. Then during the tracking procedure, the subset of the improved HOGC that most compactly discriminate the tracked object from its background is selected as the object sparse representation by calculating the L1-norm minimization based on the improved HOGC for object/sample representation. Finally, the object is tracked with the discriminative sparse representation.

The rest of the paper is organized as follows. Object tracking based on sparse representation is described in Section 2 in detail. Experiments are presented in Section 3. Section 4 concludes this paper.

Sponsored by Pri-SDL lab in GUCAS

978-1-4673-0199-2/12/$31.00 ©2012 IEEE

II. THE TRACKING METHOD

The key of object tracking is to find its location and scale s in a searching window of a new video frame. The flow chart of our tracking algorithm is shown in Fig. 1. We will discuss details of the approach in section 2.1, 2.2, 2.3, 2.4 and 2.5.

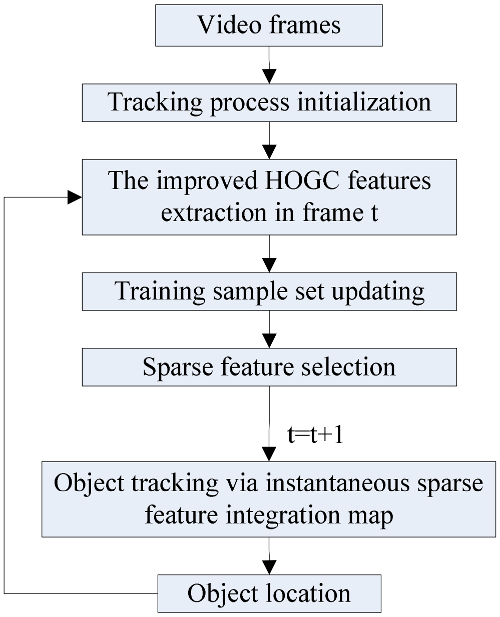

Video frames
Tracking process initialization
The improved HOGC features extraction in frame t
Training sample set updating
Sparse feature selection
t=t+1
Object tracking via instantaneous sparse feature integration map
Object location

Figure 1. Flow chart of the proposed tracking algorithm

2.1 Tracking process initialization

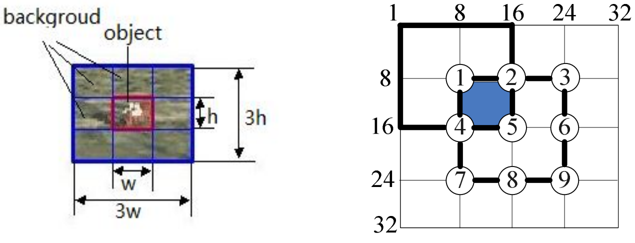
The object location and the sample set should be initialized before we can track the object in the video.

Objects like vehicles, humans and faces can be detected using corresponding object detection methods, which have been largely developed in the last few years. So objects can be located by its corresponding detection methods, or we can locate the object manually.

Similarly, the sample set can be constructed for both the object and its background by the first M tracking results (positive samples) and their corresponding backgrounds (negative samples), or we can locate the object in the first M frames to construct the training samples. In our experiments, we use the first to construct the sample set. Then we obtain $2M$ samples for the set. Then we use the improved HOGC to represent each sample in the set. The region definition and extraction of the improved combined feature are deeply described in section 2.2.

2.2 The improved HOGC feature extraction

Compared with [10], the tradition HOGC is improved by respectively extracting HC features in 9 blocks in an image window as HOG features. Details of the improved HOGC features extraction are described as follows. We define the object and background regions by rectangles as shown in Fig. 2 (a). If an object is represented by a rectangle whose area is $w \times h$ pixels, the corresponding background is then defined as 8 rectangles (each is with the same size $w \times h$ of the tracked object). For improved HOGC feature extraction, taking the object for example, first resize the object rectangle region into an image window of fixed size (32x32 pixels in our experiments); then divide the image window into small spatial regions ("cells") with the size of 8x8 pixels and each group of 2x2 cells is integrated into a block in a sliding fashion as shown in Fig. 2 (b), therefore block overlaps each other; finally 8 orientations of HOG features and 48 dimensions of HC features are extracted in each block. Therefore, in the improved combined feature set, there are totally N=504 $(9 \times (48+8))$ features corresponding to R, G and B color components in RGB color space and HOG features in gray scale space.



a. object and its background

b. Region partition

Figure 2. The improved HOGC extraction

Hence, local color information is included in the improved HOGC features, so objects with different local color but same global color histograms can be distinguished here.

2.3 Training samples updating

While in most of existing approaches, a bad template updating tends to cause the template drift problem and then leads to tracking failures, especially when there are object appearance variations or occlusions, we avoid this by updating the training samples during tracking instantaneously.

For each tracking frame, we randomly choose a positive sample and its corresponding negative sample from the set and replace it with the latest tracking result and its corresponding background. The updating of the sample set ensures that most recent object and background appearances are reflected in the sample set. It should be noted that only two samples in the set are replaced with the latest tracking result and background each time, and so even a bad sample replacement during the tracking process of the proposed updating strategy affects little on the whole set used for sparse representation, which effectively avoids the drift problem [16] and ensures tracking stability.

2.4 Sparse feature selection

In this section, L1-norm minimization based feature selection from a group of dense HOGC features, aiming to find a subset of dominant features with large discriminative ability, is used for sparse representation. The procedure is formulated as an optimization model as

$$\min \qquad \| w \|_1 \tag{1}$$

$$\text{s.t.} \quad y_i \cdot (w^T x_i) \geq \alpha \tag{2}$$

where $\|\bullet\|_1$ denotes L1-norm, (2) is the constraint to (1), ensuring that training samples will be correctly classified, $w \in R^N$ is the feature discriminative ability, $x_i \in R^N$ is the improved HOGC feature vector of the i^{th} sample, and y_i is the class label of the i^{th} sample, $y_i \in \{-1,1\}$. α guarantees that the shortest distance of different classes is 2α.

It is known that L1-norm is not differentiable, so the optimization model is difficult to be solved with a direct method as a disciplined convex program. There is, however, a simple and relatively common transformation that allows this problem to be solved effectively. Vectors are introduced here. $u \in R^N$, $v \in R^N$ and the substitution $w = u - v, u \geq 0, v \geq 0$. These relationships are satisfied by $u^j = (w^j)_+$, $v^j = (-w^j)_+$, $j = 1, 2, \ldots N$. j denotes the dimension of feature vector, and $(\bullet)_+$ denotes the positive-part operator defined as $(w^j)_+ = \max\{0, w^j\}$. Thus we get $\|w\|_1 = I_N^T u + I_N^T v$, where $I_N = [1,1,1,\ldots 1]^T$ is an N-dimension unit vector. Then we can rewrite (1) and (2) as the following disciplined convex program model:

$$\min \quad I_N^T u + I_N^T v \tag{3}$$

$$\text{s.t.} \quad \begin{cases} y_i \cdot (u-v)^T x_i \geq \alpha \\ u \geq 0 \\ v \geq 0 \end{cases} \tag{4}$$

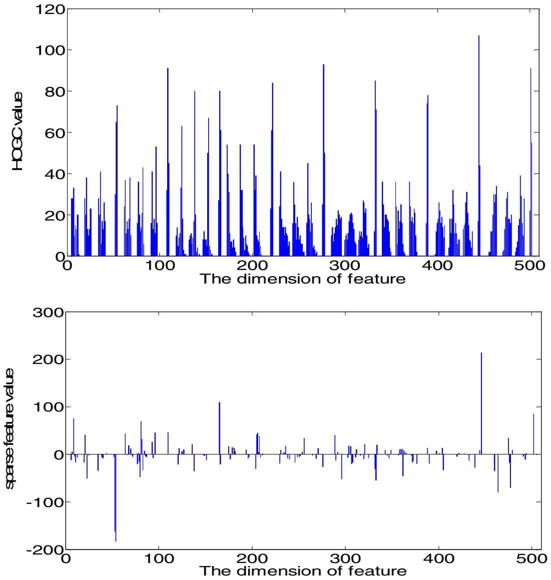


Figure 3. The comparison between sparse representation and the original HOGC representation of a positive

The optimization model shown in (3) and (4) is equivalent to (1) and (2), where u and v are two new variables of the model. The new model can be solved by Interior Point Method refer to [17]. In the case of no linear solutions are obtained, Dimensions will be firstly raised by kernel l1 minimization. The comparison between sparse representation and the original HOGC representation of a positive can be seen in Fig. 3.

2.5 Object tracking based on sparse representation

When get n features from N features by the sparse feature selection, where n<<N, then, exhaustive search method is exploited to track the object. Our goal is to search the object location in a searching area Ω_t by minimizing the difference of the n selected features between the candidates in Ω_t and the initialized object.

$$\min_{pos \in \Omega_t} (\| F(t_0) - F(C_t(pos)) \|_1) =$$

$$\min_{pos \in \Omega_t} (\| \sum_{i=1}^{n} (F(t_0)_i - F(C_t(pos)))_i \|_1) \tag{5}$$

where $F(t_0)$ is the sparse HOGC feature of the initialized object, $C_t(pos)$ is the candidate at location pos in Ω_t, $F(C_t(pos))$ represents the feature vector of candidate $C_t(pos)$, and F_i is the i^{th} selected feature of sparse representation.

Exhaustive searching method can calculate the most accurate location in local areas, but it is complex and time consuming, especially when the area of the object is large.

To get the result faster and make the tracking to be real time, integration map is employed in the candidate HOGC features calculating process. In this way, the feature arrays can be obtained by simple adds and subtracts calculates.

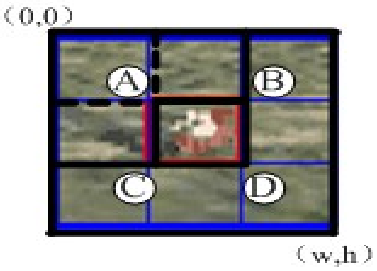


Figure 4. Integration features map to calculate the features of candidates

The features of candidates in the middle can be obtained by integration features as

$$F = J(D) - J(B) - J(C) + J(A) \tag{6}$$

where $J(\bullet)$ is the integration features from the base point $(0,0)$ to $(\bullet)$.

III. EXPERIMENTS

In this section, experiments with comparisons are carried out to validate the proposed approach. The experimental videos are from VIVID, CAVIAR and SDL data set [18].

A variety of cases are included in the test videos, such as appearance variations, scale variations, object rotations and complex backgrounds. The objects include moving humans and vehicles. Experimental results on three video clips of them are shown in Fig. 5.

In the first video clip in Fig. 5 (a), the video is captured on a moving platform and the tracked truck runs in grassland. The object has similar appearance with its background, and both of them keep changing during almost all the tracking process. Meanwhile, the object varies in size and shape during the tracking process, which increases the tracking difficulties.

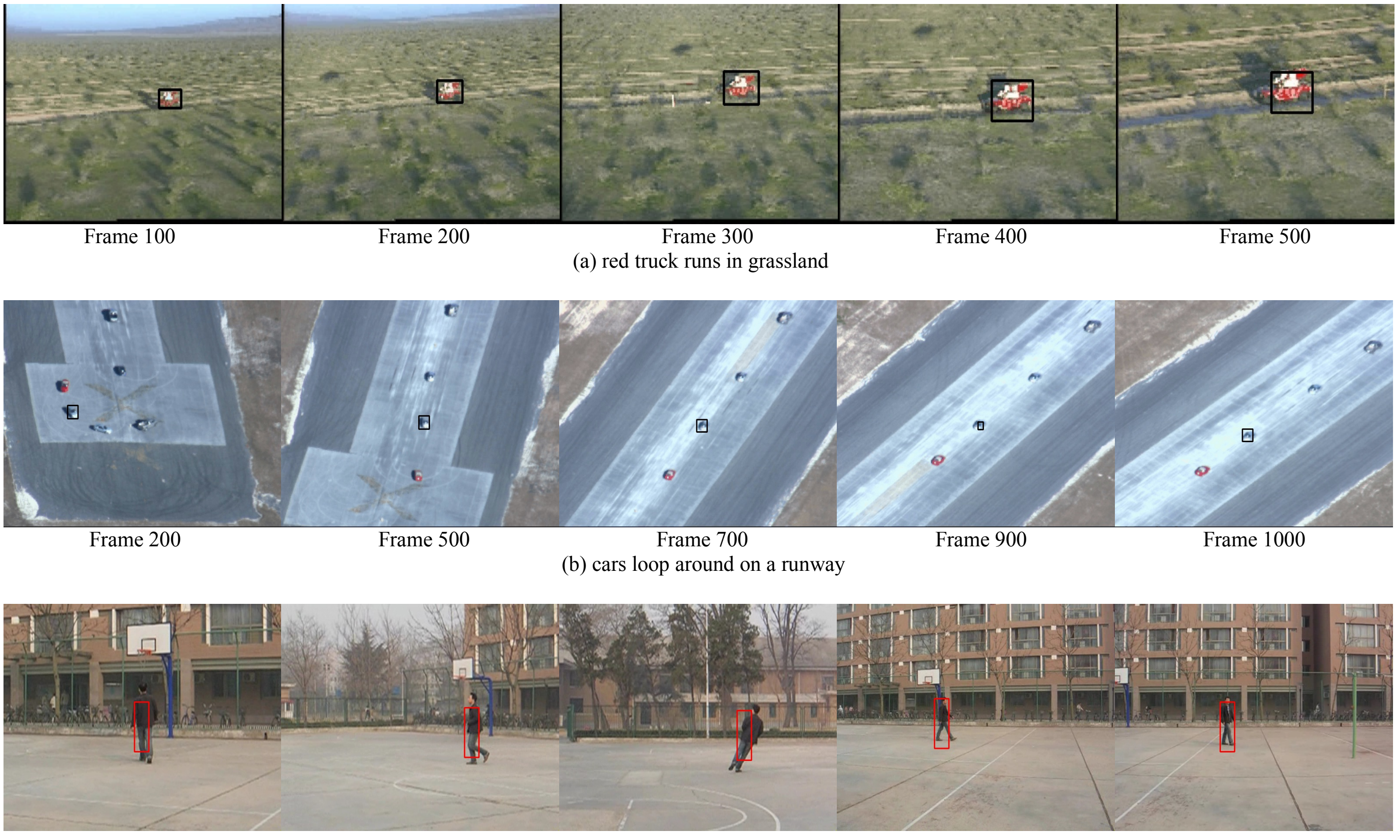
In the second video clip, the tracked car first loops around on a runway, then speeds up and runs directly. The appearances vary largely from the initial state because of rotations, and the car has similar color with the background. Tracking results of the white car are shown in Fig. 5 (b).

The third video clip shown in Fig. 5 (c) from the SDL data set is quite challenging. The man walks or runs around the basketball playground, with drastic appearance variations. There are trees and basketball stings that look like the human appearances and the man turns around sometimes so the appearances and colors are always changing during the tracking procedure. The man is tracked correctly in most of cases although the background is quite the same with humans both in appearances and colors in some frames.

In our experiments we use the DER (displacement error rates) to evaluate the tracking algorithm, which are shown in Figure 6. It can be seen that the DER (about 0.1 to 0.2) of our method is quite small in the whole tracking process. Following the idea of [7], we define the relative displacement error rate (DER) as

$$DER = \frac{d_E(T_e, T_{GT})}{\sqrt{A_{GT}}} = \frac{\sqrt{(a_e - a_{GT})^2 + (b_e - b_{GT})^2}}{\sqrt{A_{GT}}} \qquad (7)$$

where $(a,b)_e$, $(a,b)_{GT}$ and A_{GT} are the center locations of the target estimated, annotated and the size of the annotated object, respectively. We use the DERs of video clips to evaluate the proposed algorithm in Fig.6 and the statistic result verifies the efficiency of our method.



(a) red truck runs in grassland

(b) cars loop around on a runway

(c) human walks or runs in a basketball playground

Figure 5. sample frames of tracking experiments using our proposed approach

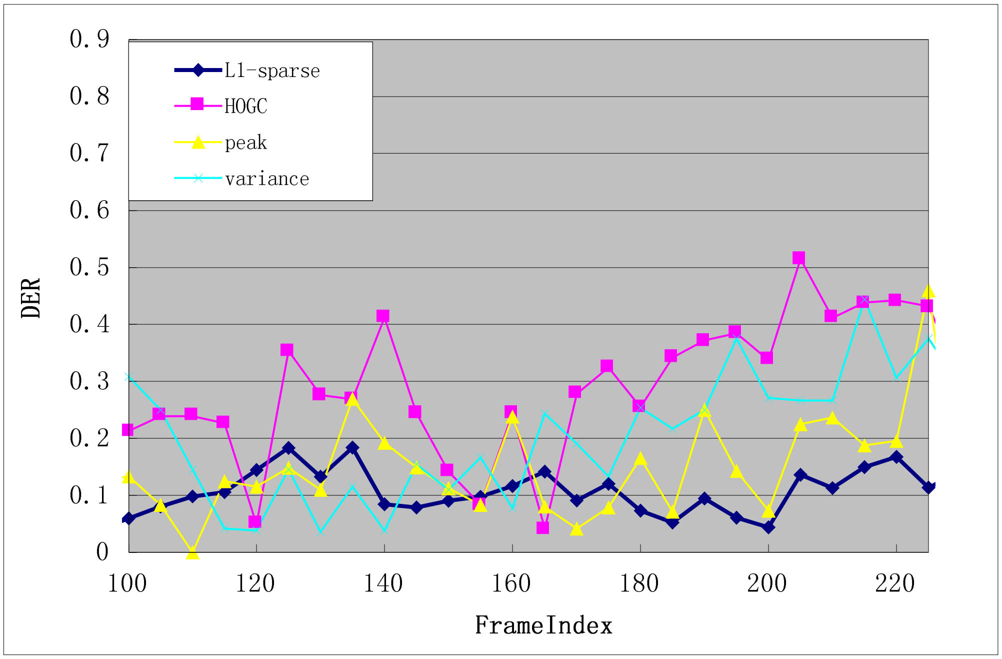


Figure 6. Displacement Error Rate (DER) of our method, HOGC (HOGC-based tracking without sparse feature representation), peak method (Peak difference feature shift) and variance method (Variance Ratio feature shift)

IV. CONCLUSIONS AND FUTURE WORKS

Object representation is very important to improve the adaptability of visual object tracking. In this paper, we have proposed a novel object tracking approach via online sparse feature selection by exploiting the L1-norm minimization based on dynamically constructed and updated sample set. The new concepts and techniques introduced in this paper include the sample set, the improved HOGC, and sparse representation based on feature selection.

A known issue in the proposed tracking method is that the whole object occlusion problem or even partial occlusion in a relatively long time has not been solved yet, for our proposed approach can not predict the position of the object in the following video frames. This issue should be considered in the future work.

ACKNOWLEDGMENT

This work is supported in part by National Basic Research Program of China (973 Program) with No. 2011CB706900, 2010CB731800, and National Science Foundation of China with No. 60872143.

REFERENCES

[1] Datong Chen and Jie Yang, "Online Learning of Region Confidences for Object Tracking, " *ICCV*, 2005.

[2] Blackman, S. and Popoli, R, "Design and Analysis of Modern Tracking Systems, " *Artech House*, 1999.

[3] Dawei Liang, Qingming Huang, Shuqiang Jiang, Hongxun Yao and Wen Gao, "Mean-Shift blob tracking with adaptive feature selection and scale adaption," *ICIP*, 2006.

[4] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust kalman filter," ICCV, 2003.

[5] S. Zhou, R. Chellapa and B. Moghadam, "Adaptive visual tracking and recognition using particle filters," Proceedings IEEE International Conference on Multimedia and Expo (ICME). 349–352, 2003.

[6] D. Comaniciu and P. Meer, "Mean shift analysis and applications," In IEEE International Conference on Computer Vision (ICCV). Vol. 2. 1197–1203, 1999.

[7] Robert T. Collins and Yanxi Liu, "Online Selection of Discriminative Tracking Features," *Proceedings of Ninth IEEE ICCV*. Vol 1, 346–352, 2003.

[8] Navneet Dalal, Bill Triggs, "Histograms of Oriented Gradients for Human Detection," *CVPR*, 2005.

[9] Zhenjun Han, Qixiang Ye, Jianbin Jiao, "Online feature evaluation for object tracking using Kalman filter," *ICPR*, 2008.

[10] Shai Avidan, "Ensemble Tracking," *IEEE Transactions on PAMI*, vol. 29(2), 261-271, 2007.

[11] Shy Shoham, Daniel H. O'Connor and Ronen Segev, "How silent is the brain: is there a "dark matter" problem in neuroscience", *Journal of Comparative Physiology*, Volume 192(8), 777-784, 2006.

[12] Mario A. T. Figueiredo, Robert D. Nowak, Stephen J. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems," *IEEE Journal of Selected Topics in Signal Processing*, 2007.

[13] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on PAMI*, 2008.

[14] Xue Mei, Haibing Ling, "Robust visual tracking using l1 Minilization," ICCV, 2009.

[15] Ran Xu, Baochang Zhang, Qixiang Ye, Jianbin Jiao, "Human detection in images via L1-Norm Minimization learning," *ICASSP*, 2010.

[16] L. Matthews, T.Ishikawa, S.Baker, "The template update problem," *IEEE Transactions on PAMI*, vol. 26(2), 810-815, 2004.

[17] Margaret H. Wright, "The interior-point revolution in optimization: history, recent developments, and lasting consequences," Bull. Amer. Math. Soc. (N.S.), Vol 42, 2005.

[18] http://coe.gucas.ac.cn/SDL-Homepage/resource.html