Online Object Tracking and Learning with Sparse Deformable Template Models

Bowen Shi [1], Tianzhe Fan [2], Qun Liu [3]
[1] Department of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu, China
[2] Jackson High School, Massillon, USA
[3] Department of Computer Science, Troy University, Troy, USA
E-mail: fergusbrilliant@hotmail.com

Abstract—**Object tracking is an important and challenging task in the field of computer vision. The objective of object tracking is to associate target objects in consecutive video frames. In this paper, we use SVM trained active basis model as a sparse deformable template for representing objects. Active basis model is a sparse model, which represents each image as a small number of bases selected from an over-complete Gabor dictionary. Given the bounding box of the object in the first frame, the model can be trained on the positive image patch inside the bounding box and negative images outside the bounding box. The tracking is achieved by detection of the object in the subsequent frames in the video by using the learned model. The model will be updated after a few of frames by using new positive and negative images, which are specified by the model. The experiment shows a good performance of the tracking method in some testing videos.**

Keywords-object tracking; support vector machoine; active basis model; deformable template; computer vision

I. INTRODUCTION

Object tracking is an important and challenging task in the field of computer vision. The availability of the high-powered computers, the high quality of inexpensive cameras, and the increasing demand for video analysis have generated a great interest in studies of the video analysis and understanding. There are three key steps in understanding a video: (1) detection of the moving object, (2) tracking of the detected object from frame to frame, and (3) analysis of the moving object to recognize their behavior. Therefore, object tracking becomes very crucial in all the video analysis problems, such as action detection, and event classification from videos [1].

The objective of object tracking is to associate target objects in consecutive video frames, which generally includes three key components: (1) object representation (2) object classifier, and (3) tracking strategy.

There are a lot of ways to represent objects. The probability density models for object appearance include Gaussian model [2], a mixture of Gaussians [3], Parzen windows [4], and histograms [5]. Templates model using simple geometric shapes to represent objects, include [6] and [7]. Additionally, there are multiple ways to train classifiers: SVM [8] and logistics regression [9]. The former one seeks to find the hyperplane that has the largest distance to the nearest training-data point of any class, while the latter explicitly models the probability of a given data belonging to a certain class.

In order to account for shape deformation of the object as well as use a sparse representation of the object, we use one type of deformable template model to represent the object in our paper. The model we used is the active basis model [7], which is a special case of sparse coding [10]. Active basis model represents the object by a small number of selected basis functions from an over-complete dictionary. It also allows the selected basis to perturb locally to account for shape deformation. The active basis model help reduce the original dimension of the objects. The shared sketch algorithm [7], which is a variant of the matching pursuit algorithm [11], is used to learn active basis model from the training images, which involves selecting a small number of bases from the dictionary. SVM with L2 regularization is used to train classifier on top of the active basis model in our paper.

The general framework we use in the object tracking is as follows: At the first step, active basis model is learned on the object patch cropped from the first frame in the video (the bounding box of the object in the first frame is given in the video tracking problem). Them we train SVM by using positive images (inside the bounding box) and negative images (outside bounding box). For the subsequent frames, we track the target by detecting them via the SVM classifier and feature responses extracted by active basis model. After a few steps of tracking, we identify more positive and negative image patches, and update both active basis model and SVM by using the new positive and negative images.

The rest of this article is organized as follows. In Section 2, we will present a sparse deformable template model, which is a SVM trained active basis model. In Section 3, we will present a framework of object tracking. In Section 4, we test our framework in some testing videos. The experiment shows good performances of the proposed method. Finally, section 5 presents a conclusion of the paper.

II. SPARSE DEFORMABLE TEMPLATE MODELS

A. An Overcomplete Dictionary of Gabor Filters

Gabor filter is a linear filter used for edge detection. Frequency and orientation representations of Gabor filters are similar to those of the human visual system. Olshausen and Field [10] have used sparse coding to learn Gabors from

natural images, which helps us to understand the V1 cells in visual cortex. The Gabor filter have been used in the field of computer vision, image processing, and data compression. The Gabor filters are translated, rotated and dilated versions of the following function:

$$G(x_1, x_2) \propto \exp\{-[(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2]/2\} e^{ix_1}$$

which is sine-cosine wave multiplied by a Gaussian function. The Gaussian function is elongated along the x_2-axis, with $\sigma_2 > \sigma_1$, and the sine–cosine wave propagates along the shorter x_1-axis. We then translate, rotate and dilate $G(x_1, x_2)$ to obtain a general form of the Gabor wavelets:

$$B_{x_1,x_2,s,\alpha}(x_1', x_2') = G(\tilde{x}_1/s, \tilde{x}_2/s)/s^2,$$

where

$$\tilde{x}_1 = (x_1' - x_1)\cos\alpha + (x_2' - x_2)\sin\alpha,$$
$$\tilde{x}_2 = -(x_1' - x_1)\sin\alpha + (x_2' - x_2)\cos\alpha.$$

For simplicity, we use $x = (x_1, x_2)$ to denote 2D coordinate for locations, each $B_{x,s,\alpha}$ is a localized function, where x is the central location of the Gabor, s is the scale, and α is the orientation. We discretize s and α into finite proper ranges. The dictionary of Gabor filters is denoted by $\Omega = \{B_{x,s,\alpha}, \forall(x, s, \alpha)\}$. Figure 1 visualizes some Gabor filters at different locations, orientations, and scales.

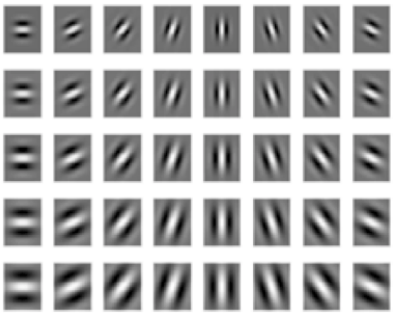

Figure 1. Visualization of Gabor filters at different locations and scales

B. Active Basis Model as a Composition of Gabors

Suppose we have a dictionary of Gabor wavelets $\Omega = \{B_{x,s,\alpha}, \forall(x, s, \alpha)\}$. We denote $\{I_m$, m = 1, ..., M$\}$ as a set of training images defined on a common rectangle lattice D, and suppose that these images are from the same object category and are roughly aligned. According to sparse coding [10], each image can be represented by a small number of Gabor filters selected from the dictionary:

$$I = \sum_{i=1}^{n} c_i B_{x_i,s_i,\alpha_i} + \epsilon, \quad (1)$$

where c is the projection coefficiens, and ϵ is the reconstruction error. It give us a sparse representation of an image. Active Basis model [7] proposed share sparse coding to represent a group of M training images simultaneously by a common set of selected Gabors:

$$I_m = \sum_{i=1}^{n} c_{m,i} B_{x_i,s_i,\alpha_i} + \epsilon_m, \ \ m = 1, \dots, M \quad (2)$$

In order to account for shape deformation that appears in the images, Active basis model [7] further introduces the perturbations into the selected Gabors, and the model can be defined as

$$I_m = \sum_{i=1}^{n} c_{m,i} B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}} + \epsilon_m, \ \ m = 1, \dots, M \quad (3)$$

$\mathbf{B} = (B_{x_i,s,\alpha_i}, i = 1, \dots, n)$ can be considered a common template for the training images, which is a composition of Gabor filters. This template is deformable. For each image I_m, the Gabor wavelet element B_{x_i,s,α_i} is perturbed to $B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}$ by local maximum, where $\Delta x_{m,i}$ is the perturbation in location, and $\Delta\alpha_{m,i}$ is the perturbation in orientation. $\mathbf{B}_m = (B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}, i = 1, \dots, n)$ can be considered the deformed templates for coding image $\mathbf{I}_m$.

C. Shared Sketch Algorithm for Basis Selection

Matching pursuit is a greedy way to select bases for encoding a given image, which is used to solve the model in (1). For dealing with model in (3), shared sketch algorithm [10] is proposed for learning active basis model. We present a pseudo code as follows:

(0) Initialize $i \leftarrow 0$. For $m = 1, \dots, M$, initialize $R_m(x, \alpha) \leftarrow \langle I_m, B_{x,s,\alpha} \rangle$ for all (x, α).

(1) $i \leftarrow i + 1$. Select

$$(x_i, \alpha_i) = \arg\max_{x,\alpha} \sum_{m=1}^{M} \max_{(\Delta x, \Delta\alpha) \in A(\alpha)} h(|R_m(x + \Delta x, \alpha + \Delta\alpha)|^2).$$

(2) For $m = 1, \dots, M$, retrieve

$$(\Delta x_{m,i}, \Delta\alpha_{m,i}) = \arg\max_{(\Delta x, \Delta\alpha) \in A(\alpha)} |R_m(x_i + \Delta x, \alpha_i + \Delta\alpha)|^2.$$

Let $c_{m,i} \leftarrow R_m(x_i + \Delta x_{m,i}, \alpha_i + \Delta\alpha_{m,i})$, and update $R_m(x, \alpha) \leftarrow 0$ if $\text{corr}\left(B_{x,s,\alpha}, B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}\right) > 0$.

(3) Stop if $i = n$, else go back to 1.

$h()$ is a sigmoid function, which increases from 0 to a saturation level ξ (default: $\xi = 6$),

$$h(r) = \xi\left[\frac{2}{1 + e^{-2r/\xi}} - 1\right].$$

And $\langle I_m, B_{x,s,\alpha} \rangle$ is a filtering operation that captures some image features. $\text{corr}\left(B_{x,s,\alpha}, B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}\right)$ is the correlation between filters $B_{x,s,\alpha}$ and $B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}$.

D. Discriminative learning by SVM for Object detection

After selecting the bases, we can use them to extract features from the images. In particular, for the basis template

$\mathbf{B} = (B_{x_i,s,\alpha_i}, i = 1,\ldots,n)$, we can extract feature vector $(X_1, X_2, \ldots, X_n)$ from each image by deforming the template on the image so that the total template response is maximized. After extracting features, support vector machine (SVM) [8] can be used to train a discriminative classifier on these feature vectors extracted by the selected bases. SVM is s supervised learning algorithm, which need negative examples for training.

SVM benefits from two good ideas: maximizing the margin and the kernel trick. SVM seeks to construct a hyper-plane as the decision surface, which maximize the margin of separation between positive and negative examples. In addition, kernel functions are widely used in SVM to deal with non-linear classification problems. The kernel function maps the original input data into a higher dimensional space so that it can construct the optimal hyper-plane to separate the data. Some popular kernel functions include the polynomial, radial based and the sigmoid functions. The output of SVM classification is the decision values of each instance for each class, which are used for probability estimates. The probability values represent "true" probability in the sense that each probability falls in the range of 0 to 1, and the sum of these values for each instance equals 1. Classification is then performed by selecting the highest probability [8] [12].

The above algorithm is good for classification, which means that the objects in the training images and the testing images are roughly aligned. As to detection, in which we need to localize the object in a testing image by indicating the locations of the objects. Usually we place a bounding box on the detected object to show the detection result. Sliding window is a standard technique for object detection with a training classifier. To detect the object in a new image, we scan the classifier or the template over the testing image to get probability scores in all locations, also we allow the bases, that is, Gabor filters locally move to account for deformation within a finite proper range during the process of sliding windows. In addition, to achieve multi-scale detection, we incrementally resize the testing image and perform sliding window algorithm over each of these resized images, which is equivalent to using different scales of template to detect the object within a fixed size of image. This process is efficient with dynamic programming, which gives us a bottom up and top down computational framework.

III. OBJECT TRACKING BY LEARNING

Given a specified object in the first frame of a video, the goal of online object tracking is to locate the object in the subsequent frames by placing bounding boxes on it. We will present a tracking by learning framework based on the deformable template model, which include the following three steps:

(1) Model learning from the object given in the first frame of the input video. In this step, we firstly learn an active basis model from the object by shared sketch algorithm, and then use the learned template to extract features from both positive and negative images. The positive images are the image patch inside the given bounding box, while the negative images are the image patches outside the given bounding box. A SVM is trained on those extracted features.

(2) Object tracking by detection. After we learn the model based on the first frame, we can use it to track the object in the subsequent frames by detection. The detection algorithm has been mentioned at the previous section. However, instead of performing sliding window over the whole image, we can just scan the small region locally around the bounding box provided in the previous frame.

(3) Update the model by relearning. After we locate the target, we will have more training images to relearn the model so that we will obtain a more precise and correct model for the subsequent tracking.

We call it tracking by learning. Figure 2 illustrates the basic idea of the online object tracking framework.

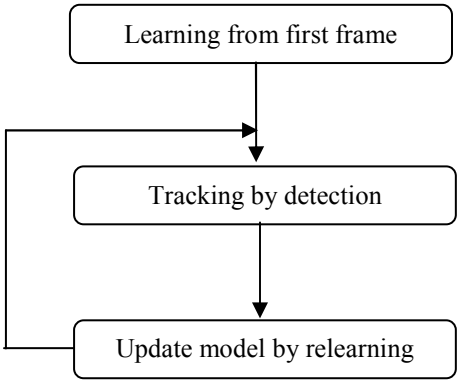

Figure 2. Framework of object Trsacking

IV. EXPERIMENTS

To verify the effectiveness of the tracking framework, we test the tracking method in some videos collected from KTH video dataset [13]. In each video, the bounding box in the first frame is assumed to be given and known to the algorithm. We learned an active basis model from the image patch cropped from the first frame based on the given bounding box. After we select the bases by shared sketch algorithm, then we train a SVM classifier on this sparse model by using the positive patch inside the bounding box and the negative patches outside the bounding box. We use the learned model to track the objects in the subsequent frames by locally detecting the objects. After we locate the objects, then we collect new positive and negative images in the subsequent frames based on the detection results, then we update the SVM classifier by relearning the active basis model and re-train the SVM parameters. The number of selected basis is 70, and location and orientation perturbation for the active basis model are 2 pixels and 2 levels of orientations (each level of orientation is pi/16) respectively. Figure 3 shows some tracking results on four videos. We place the bounding box on each frame in the video to illustrate the tracking results. We can see that the tracking method works well in these testing videos.

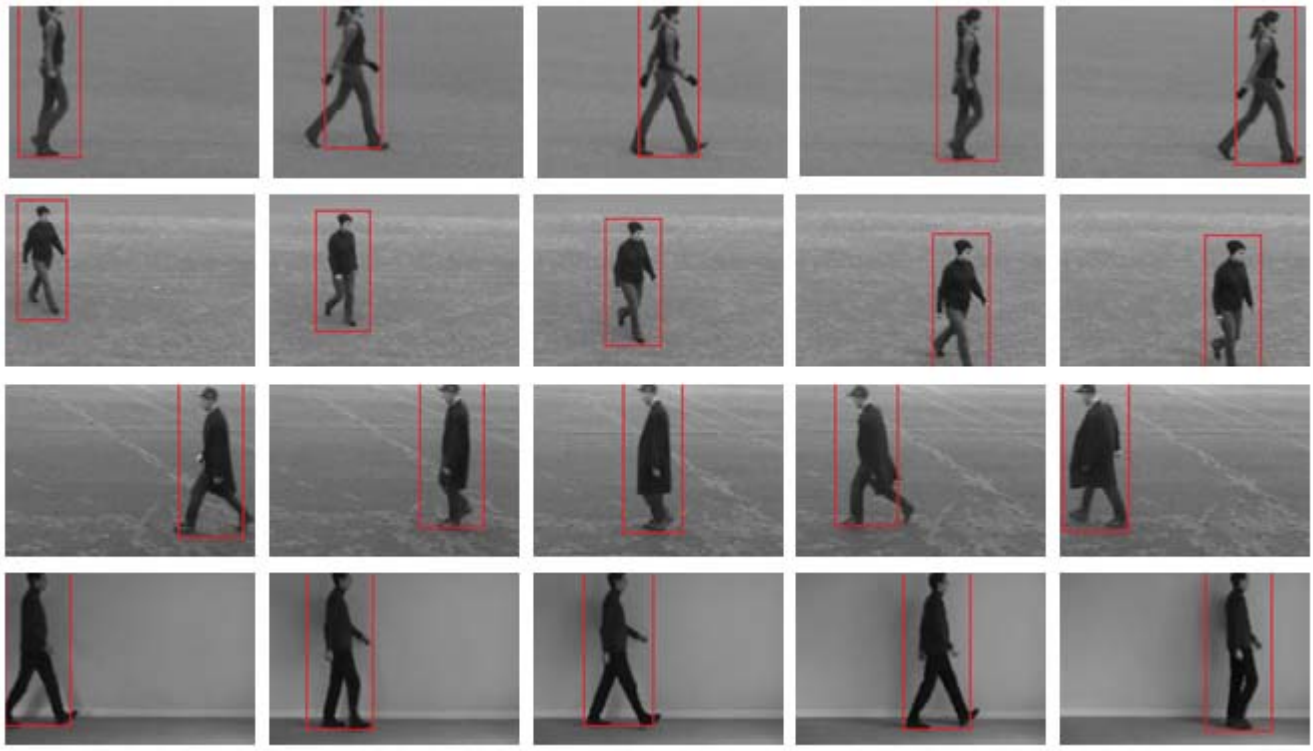

Figure 3. Object tracking. Each row illustrates a result of the object tracking in a video. The bounding box in the first frame is given. The bounding boxes shown in the subsequent frames are predicted by the learned model.

V. CONCLUSIONS

In this paper, we present a framework of object tracking by using sparse deformable template model as well as SVM classifier. This tracking by learning method performs object tracking by detection, and simultaneously update the model using the new detected objects. Because there are variability between objects in different frames in a video, we apply the sparse deformable template model to account for shape deformations. To better separate the object and the background robustly, SVM is used for discriminative learning. The experiments show good performance of the proposed method in tracking and detecting objects in videos.

REFERENCES

[1] Yilmaz, Alper, Omar Javed, and Mubarak Shah. "Object tracking: A survey." Acm computing surveys (CSUR) 38.4 (2006): 13.

[2] Zhu, Song Chun, and Alan Yuille. "Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 18.9 (1996): 884-900.

[3] Paragios, Nikos, and Rachid Deriche. "Geodesic active regions and level set methods for supervised texture segmentation." International Journal of Computer Vision 46.3 (2002): 223-247.

[4] Elgammal, Ahmed, et al. "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance." Proceedings of the IEEE 90.7 (2002): 1151-1163.

[5] Comaniciu, Dorin, Visvanathan Ramesh, and Peter Meer. "Kernel-based object tracking." Pattern Analysis and Machine Intelligence, IEEE Transactions on 25.5 (2003): 564-577.

[6] Fieguth, Paul, and Demctri Terzopoulos. "Color-based tracking of heads and other mobile objects at video frame rates." Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on. IEEE, 1997.

[7] Wu, Y. N., Si, Z., Gong, H., & Zhu, S. C. (2010). Learning active basis model for object detection and recognition. International journal of computer vision, 90(2), 198-235.

[8] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[9] Hosmer Jr, David W., and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.

[10] Olshausen, Bruno A., and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?." *Vision research* 37.23 (1997): 3311-3325.

[11] Tropp, Joel, and Anna C. Gilbert. "Signal recovery from random measurements via orthogonal matching pursuit." Information Theory, IEEE Transactions on 53.12 (2007): 4655-4666.

[12] Beniwal, Sunita, and Jitender Arora. "Classification and feature selection techniques in data mining." International Journal of Engineering Research & Technology (IJERT) 1.6 (2012).

[13] Laptev, Ivan, and Tony Lindeberg. "Velocity adaptation of space-time interest points." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 1. IEEE, 2004.