A Generic Approach to Object Matching and Tracking

Xiaokun Li[1], Chiman Kwan[1], Gang Mei[1],
and Baoxin Li[2]

[1] Signal/Image Processing Group, Intelligent Automation Inc.,
Rockville, MD 20855, USA
[2] Computer Science and Engineering Dept., Arizona State University,
Tempe, AZ 85287, USA

Abstract. In this paper, a generic approach to object matching and fast tracking in video and image sequence is presented. The approach first uses Gabor filters to extract flexible and reliable features as the basis of object matching and tracking. Then, a modified Elastic Graph Matching method is proposed for accurate object matching. A novel method based on posterior probability density estimation through sequential Monte Carlo method, called as Sequential Importance Sampling (SIS) method, is also developed to track multiple objects simultaneously. Several applications of our proposed approach are given for performance evaluation, which includes moving target tracking, stereo (3D) imaging, and camera stabilization. The experimental results demonstrated the efficacy of the approach which can also be applied to many other military and civilian applications, such as moving target verification and tracking, visual surveillance of public transportation, country border control, battlefield inspection and analysis, etc.

Keywords: Image analysis, feature extraction, object matching, real-time tracking.

1 Introduction

In image analysis and recognition, automatically recognizing objects of interest is always a challenging problem, and has been a research topic for many years. In recent years, detecting and tracking moving object in video is becoming a more interesting research topic and alluring more research efforts. In low level computer vision, one fundamental problem in object recognition and tracking is feature extraction as the result of extraction will directly affect the recognition performance. Another tough problem in object recognition is the matching between target and template. One reason for these difficulties is that, in real world, the object of interest always has some orientation difference and shape deformation as compared to its template in database. The goal of this paper is to develop an efficient method for object recognition and verification. The proposed method is based on Gabor filter-based Elastic Graph Matching (EGM) which has been successfully used in image texture analysis, face and fingerprint recognition [1-5]. But, by applying a new template-based matching method as the initialization of EGM, which is invariant to object rotation and size, we can overcome the limitations of conventional EGM and extend its applicability to more general cases such as stereo imaging, object tracking, and image sequence stabilization.

A. Campilho and M. Kamel (Eds.): ICIAR 2006, LNCS 4141, pp. 839 – 849, 2006.

Another important issue discussed in this paper is object tracking. In video/image analysis, object tracking often becomes a more desirable problem after recognition. An automatic algorithm is needed to answer the questions: what is trace of the detected object? Or, is the object in the current frame the one I am looking for? Once the object is detected, people usually want to know its status and position in the subsequent frames. In the real world, the object of interest is moving in 3D space, meaning the features of the object, which are projected onto 2D image, are also changing along the temporal axis. This makes object tracking a very challenging problem. Even with the difficulties mentioned above, many new methods and exciting results have been obtained in recent years, e.g. [6-12]. Unlike the current methods, we propose to use Sequential Importance Sampling (SIS) method to track moving object in real-time, which has several important advantages in object tracking. First, SIS is based on posterior probability density estimation through sequential Monte Carlo (MC) method. The samples used for tracking are weighted properly via MC and updated with current observation while keeping track of a slowly varying change. Second, with SIS, tacking can be completed simultaneously by using the estimated posterior density.

In this paper, a generic approach for object matching and tracking is presented. The approach consists of three steps. The first step is Gabor filter-based feature extraction which provides an efficient way for selecting object features. The second step is an improved Elastic Graph Matching for object matching. The last step is a novel approach to simultaneously tracking multiple object in video/image sequence. Our method is based on posterior probability density estimation through sequential Monte Carlo methods.

The paper is organized as follows. Section 2 gives out the technical details on how object feature extraction is formulated with Gabor filter. Section 3 describes our approach on using EGM for object matching. In Section 4, one efficient solution for fast tracking is presented. Section 5 provides some experimental results both in video and image sequence.

2 Gabor Filter-Based Feature Extraction

In human visual system (HSV), research has shown that people are sensitive to both specific orientation and spatial frequencies of object of interest. For feature representation and extraction, wavelets are good at representing orientation and frequency characteristic of object of interest. A Gabor filter bank can act as a simple form of wavelet filter bank. Because of its simplicity and optimum joint spatial/spatial-frequency localization, Gabor filter has attracted many research efforts [4-5, 13-19] and has been applied in many image analysis and computer vision-based applications, e.g. face and fingerprint analysis and recognition [1-3].

Gabor filter bank is a group of 2-D filters which record the optimal jointed localization properties of region of interest in both spatial and spectral domain. Typically, an image is filtered with a set of Gabor filters which have different or preferred orientations and spatial frequencies. To be specific, an image $I(\vec{x})$ is filtered with a set of Gabor wavelets as follows,

$$(wI)(\vec{k},\vec{x}_0) = \int \phi_{\vec{k}}(\vec{x}_0 - \vec{x}) I(\vec{x}) d\vec{x} \tag{1}$$

where $\phi_{\vec{k}}$ is the Gabor wavelet (filter) defined by

$$\phi_{\vec{k}}(\vec{x}) = \frac{\vec{k}}{\sigma^2} \exp(-\frac{\vec{k}^2 \vec{x}^2}{2\sigma^2})[\exp(i\vec{k}\vec{x}) - \exp(-\frac{\sigma^2}{2})] \tag{2}$$

with $\vec{k} = k_v e^{i\phi_\mu}$ controlling the orientation and the scale of the filters. By varying v and μ, we can get different Gabor filters with different orientations and scales. In our implementation, μ controls the orientation and is assigned by any value of 0, 1, 2, to 7 and v controls the spatial frequency and is assigned from 0, 1, and 2 with $k_v = (\pi/2)/\sqrt{2^v}$ and $\phi_\mu = (\mu\pi)/8$. After filtering with a set of Gabor filters (24 filters from the above choice of v and μ), the outputs on each pixel in the image form a 24-dimensional vector called "jet". The amplitude of the jet represents whether a pixel has significant gradient value in both orientation and frequency. Thus, it can be used to determine if this pixel is a good feature for object matching and tracking.

3 Matching

In order to correspond two images from two different sensors, called as image level matching, or find the correspondence from target to template of reference image/database for object recognition and verification, called as object level matching, we have to solve the feature correspondence (matching) problem. With the feature points detected in the previous section, we propose to use an improved Elastic Graph Matching method to solve the matching by finding the corresponding features in the target frames. Some more detailed description of EGM can be found in [5, 20]. In most cases, due to the possible arbitrary relative positioning of the sensors with different field of view (FOV), conventional EGM method may never converge to the correct position because of the position, orientation, and scale difference between target and template, and thus we propose a coarse-to-fine strategy to perform robust matching. We first roughly match the two images (target image and template image) or find the object of interest in target image by searching with template, and then use EGM method to tune the matching result. The template matching with unknown rotation and size can be formulated using a non-orthogonal image expansion approach [21]. In this method, image or object of interest will be recovered by a delta function at the template location. The convolution equation can be expressed as:

$$g(r) = f(r;\theta^0) * \delta(r - \bar{r}^0) + n(r) \tag{3}$$

where the position vector is $r^T = [r_x, r_y]$, * is 2-D convolution, and f is the template (image or object of interest) at $\bar{r}^0$. The orientation and size differences between target and template are represented by a vector θ (where θ^0 is the true parameter set).

$$\theta^T = [s, \phi] \tag{4}$$

where s is the size and ϕ is the rotation, a rotated and resized template can be given as

$$f(r;\theta) = f(\frac{1}{s}M(\phi)r) \tag{5}$$

where $M(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$. In this coarse step, maximum likelihood (ML) can be used to estimate the parameter set θ and use delta restoration method [22] for location estimation $\hat{r}$. The cost function of ML can be described as

$$l(\theta, r \mid g) = \left\| f(r;\theta) * \delta(r - \bar{r}) - g(r) \right\|^2 \tag{6}$$

The maximum likelihood solution is then obtained by minimizing Eq. (6) as

$$\{\hat{r}, \hat{\theta}\} = \operatorname{argmin} l(\theta, r \mid g) \tag{7}$$

To solve the optimization problem, a Linear Least Square Estimate (LLSE) of the delta function can be considered to use. More details can be found in [22].

Even with the method mentioned above, two images or two objects may never be able to correlate with each other exactly due to local structural and depth variations. However, this is addressed naturally by the elasticity of the matching graph in the algorithm. In this paper, we present one improved EGM method which uses Gabor jets as inputs. Its main steps are given as follows:

Algorithm: Enhanced Elastic Graph Matching (EGM)

Step 1: Find approximate position: We use the novel template matching with unknown rotation and size parameter to identify the initial correspondence/matching between target and template of reference image/database. From the correspondences, some corresponding pairs of pixels from target and template are selected as features whose magnitudes of the jets are obviously larger than that of other pixels.

Step 2: Verify position: We first average the magnitudes of the jets of each feature point. The jets to each pixel are termed as "bunch". Then, we assign the average value to the processed bunch and compute the similarity function S_a without phase comparison.

$$S_a(J, J') = \frac{\sum_j a_j a_j'}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}}$$

where a_j is the average value of the jth bunch. Alternatively, we can compute the similarity function S_ϕ with phase.

$$S_\phi(J, J') \approx \frac{\sum_j a_j a_j' \cos(\phi_j - \phi_j')^2]}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}}$$

If the similarity is larger than the predefined threshold, the result by template matching is acceptable. Otherwise, error message will be generated and the EGM process is stopped.

Step 3: Refine position and size: To the current bunch graph, we vary its position and size to tune the correspondence. For each bunch, check the four different pixels (± 3, ± 3)displaced from its corresponded position in the target image. At each position, we check two different sizes with a factor of 1.2 smaller or larger the bunch graph.

Step 4: Refine aspect ratio: A similar relaxation process as described in Step 3 is performed. But at this time, we apply the operation only to x and y dimensions independently.

4 Tracking

In the previous section, we discuss the feature correspondence between target and template, or two input images, or an image pair of two video sequences. When people want to know the status of object of interest in a single image/video sequence, target tracking becomes an interesting research topic. Since object is located in 3D space and projected onto 2D image, some features of the object will appear and some will disappear when target is moving or sensor is moving. This is an inevitable challenge facing any conventional method of feature tracking.

Under a weak perspective camera model, the motion of a planar rigid object can be approximated by a 2D affine group. Although the set of jets is defined on an object of interest, e.g. human face, which is definitely not an ideal planar object. But, if deformation of each feature point is allowed, one can still get a good approximation to the jet motions. Therefore, we model the jet motions as a 2D affine transformation plus a local deformation. We also assume the motion change between two subsequent frames is small. Unlike conventional methods, we propose to use Markov chain Monte Carlo techniques [23] for tacking. Specifically, the Sequential Importance Sampling (SIS) algorithm is used as motion predictor to find the correspondence features in two subsequent frames (*t* frame and *t+1* frame) and on-line select features by updating new weights. In the SIS approach, object motion is formulated as the evaluation of the conditional probability density $p(X_t \mid Z_t)$. At time *t*, $p(X_t \mid Z_t)$ is approximated by a set of its samples, and each sample is associated with a weight reflecting its significance in representing the underlying density (importance sampling). The basic steps of SIS are given as follows:

SIS Algorithm

Let $S_t = \left\{X_t^{(j)}, j = 1,...,m\}\right\}$ denote a set of random draws that are properly weighted by the set of weights $W_t = \{w_t^{(j)}, j = 1,...,m\}$ with respect to the distribution π_t. At each time step *t*,

Step 1. Random draw $x_{t+1}^{(j)}$ from the distribution $g_{t+1}(x_{t+1} \mid x_t^{(j)})$;

Step 2. Compute

$$u_{t+1}^{(j)} = \frac{\pi_{t+1}(x_{t+1}^{(j)})}{\pi_t(x_t^{(j)}) g_{t+1}(x_{t+1}^{(j)} \mid x_t^{(j)})}$$

$$w_{t+1}^{(j)} = u_{t+1}^{(j)} w_t^{(j)} .$$

Then $(x_{t+1}^{(j)}, w_{t+1}^{(j)})$ is a properly weighted sample of π_{t+1}.

In this algorithm, $g(\cdot)$ is called the trial distribution or proposal distribution and computed by $g_{t+1}(X_{t+1} \mid X_t) = \frac{1}{\sqrt{2\pi\sigma_1}} \exp\left\{ -\frac{(x_{t+1} - x_t)^2}{\sigma_1^2} \right\}$. Thus, the SIS can be applied recursively for *t=1,2, …*, to accommodate an ever-changing dynamical system

---

In our tracking algorithm, after obtaining a predicted $x_{t+1}^{(j)}$, we check it with the measured value in *t+1* frame. Based on the measured feature points from the frame at *t+1*, a matching error is computed for the mapped set and the measured set. According to the matching error, $u_{t+1}^{(j)}$ is computed and $w_{t+1}^{(j)}$ is then updated. Note that we do not specify any uncertainty model for individual feature points, which may be too complex to be modeled by a simple function, since it needs to account for inaccuracies in 2D approximation, uncertainty due to noise, non-rigidity of object of interest, etc. In our method, the local deformation at each jet is used to account for these factors.

Another issue during our implementation is we reduce the motion parameter space from 2-dimension (x and y directions) to one-dimension (θ). Here, we can consider a rigid object subject to motion which can be modeled by a transformation *f* parameterized by a parameter vector θ. Let X_0 denote an original parameterization of the object. X_0 can be a set of jets. Let $X = f(\theta, X_0)$ denote the transformation of X_0 into X. Under a small and continuous motion assumption, X would be similar to X_0 and θ would be very close to θ_0.

$$\begin{aligned} X &= f(\theta, X_0) = f(\theta_0, X_0) + J_\theta(\theta - \theta_0) + o(\cdot) \\ &\approx X_0 + J_\theta(\theta_0)(\theta - \theta_0) \end{aligned} \tag{8}$$

where $o(\cdot)$ denotes higher-order terms and $J_0(\cdot)$ is the Jacobian matrix with respect to θ. Consider the 2-D affine motion $f(\theta,\cdot)$ as

$$f(\theta,\cdot) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} (\cdot) + \begin{pmatrix} T_x \\ T_y \end{pmatrix}$$

Vector $\{a_{11},...,a_{22}\}$ represents 2D affine rotation and $\{T_x, T_y\}$ represents translation.

We can compute the Jacobian matrix using

$$J_\theta(\theta_0) = \left|\frac{\partial X}{\partial \theta}\right|_{\theta_0} = \begin{bmatrix} x_0 & y_0 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_0 & y_0 & 0 & 1 \end{bmatrix} \tag{9}$$

In our tracking algorithm, we take $\Delta\theta$ as X and use X to find the new position by computing the Jacobian matrix.

Algorithm: SIS-based Tracking

Initialization: The relative camera/sensor motion with a transformation group is modeled first. The motion parameters constitute a state vector distributed according to a density function $\pi(t)$. Then, we track the evolution of $\pi(t)$ over time t using the SIS algorithm, by which $\pi(t)$ is represented by a set of samples $x(t,j)$ with proper weights $w(t,j)$, $j=1,2,\ldots$

Step 1. Find a set of feature points of object of interest in the first frame ($t=0$)**.**

Step 2. For time $t>0$, track the set of (feature) points from t to $t+1$ by performing the following:

a) Each sample $x(t+1,j)$ of $\pi(t+1)$ is used to map/predict the set of feature points to time $t+1$.

b) Based on the measured feature points from the frame at time $t+1$, a matching error is computed from the mapped set.

c) The matching error is used to control the updating of $w(t,j)$ to get $w(t+1,j)$.

Step 3. At time $t+1$, we compute the expectation (the weighted mean) of the samples $x(t+1,j)$ to get the motion at that moment, and the points corresponding to the mean give the feature set at that time.

Step 4. For the frame at time $t+1$, use the EGM algorithm with small elasticity to fine-tune the result.

5 Experiments and Applications

Many tests on image/video sequences have been performed with our proposed algorithm. In this section, three tests are selected to illustrate the efficiency and possible applications of the algorithm.

5.1 Dancer Matching and Tracking

The test data sets were acquired from public domain (the website of Microsoft Research Labs). The dynamic scenes were captured by eight cameras at different views with a common synchronization control. The data set from each camera consists of 100 images (24-bits true color, 1024 x 768) at 3 frames per second. We performed two different tests on them. One is finding feature correspondence between two images acquired from two different cameras. Another one is target tracking in a single video stream.

Fig. 1 shows the feature correspondence results of two images captured from two cameras with different view points. The pixels are correctly corresponded.

Fig. 2 shows the tracking of a dancer by using video sequence from a single camera. Again our algorithm worked very well and we were able to track the dancer even though her movements were very drastic.

- **Matching**

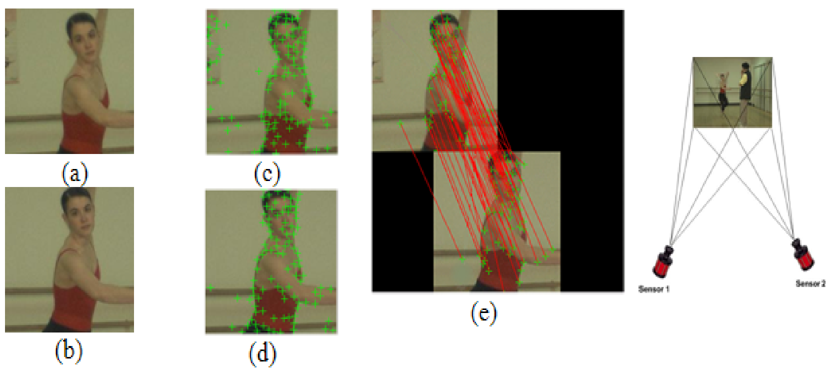

Fig. 1. Finding feature correspondence from two different-view images. (a) and (b) are two different-view images captured at the same time. (c) and (d) show the extracted feature points. (e) is the result of feature correspondence of (c) and (d).

- **Tracking**

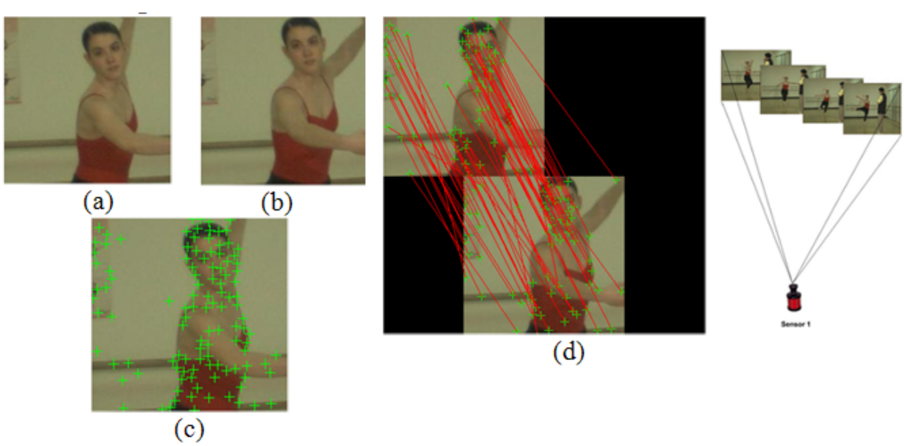

Fig. 2. Object tracking: (a) and (b) are two subsequent images acquired from same camera at different time instances. (c) shows the selected feature points from (a). The tracking result is given in (d).

5.2 Stereo (3D) Image Generation

One important application of image matching is stereo imaging. After finding the feature correspondence between two different-view images, we can use the theories of multi-view geometry [24] to generate stereo image. The testing video sets for stereo imaging were collected by two video cameras with the resolution of 640 x 480 and the frame rate of 10 f/s. We first find the correspondence of the two first frames of the two input video to create a stereo image pair. Then, the features of next stereo pairs for correspondence were tracked by using our proposed tracking method in each video stream independently. Fig. 3 shows the corresponded images and the stereo image.

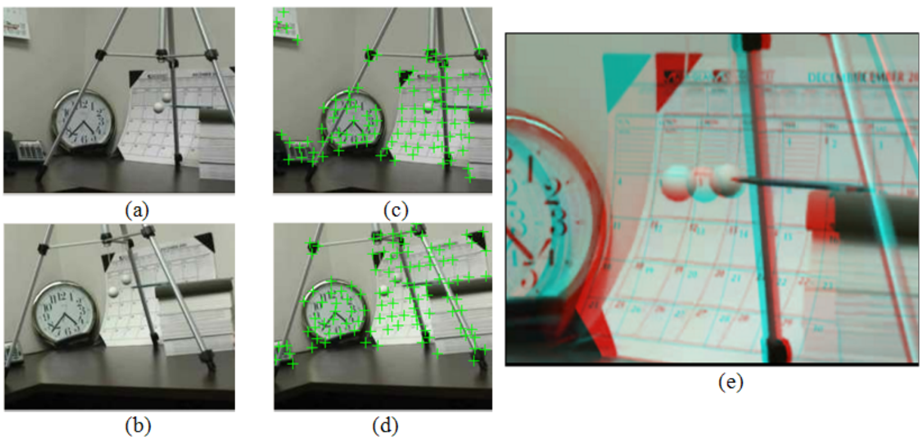

Fig. 3. Stereo imaging: (a) and (b) are two different-view images. (c) and (d) display the selected feature points for feature correspondence. (e) shows a red-cyan stereo image created from the feature correspondence (Reader can watch the 3D image with any red-cyan 3D glasses).

5.3 Target Tracking and Stabilization

One experiment was performed to show how the proposed algorithm can be applied to small target tracking and image sequence stabilization (also known as sensor motion

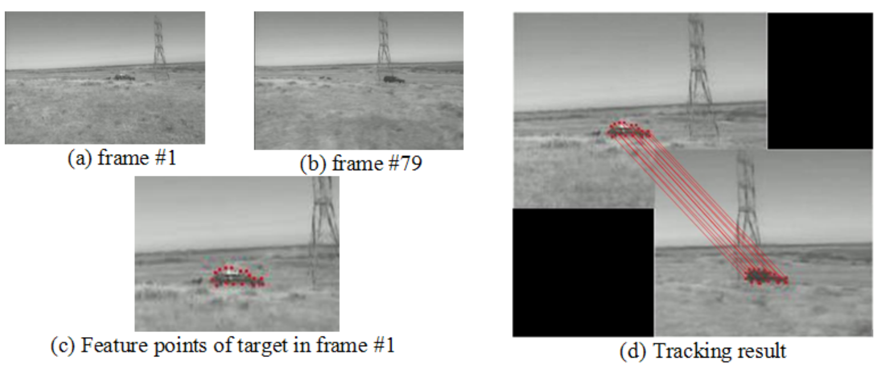

Fig. 4. Target tracking

compensation). The testing data has the resolution of 336 x 244 and the frame rate of 30 f/s. In the test, we first manually located a target of interest, e.g. a moving vehicle shown in Fig. 4 (a). Then, Gabor filter-based feature extraction and SIS-based tracking algorithms were performed to track the moving target in image sequence. The result is given in Fig. 4 (d). As we can see from the result, the small moving target can be successfully tracked in cluttered environment.

For sensor motion compensation, we modeled the camera motion as a 2D affine transformation. Stabilization is then achieved in the following steps. First, we extracted a set of feature points from the first frame. Next, we used the algorithm to track the feature set in the sequence. Stabilization was then done by warping the current frame with respect to the reference frame (the first frame) using the estimated motion parameter.

6 Summary

In this work, we have examined the problem of target matching and tracking in video/image sequence including the data acquired in noisy environment. We proposed a Gabor attribute matching and tracking algorithm based on an improved EMG and statistical sampling method. As described in the introduction section, there are many methods for object matching and tracking. Our algorithm differs from other matching methods in that we use Gabor attribute as features and extend the typical EMG method by introducing an efficient template matching method. . Consequently, our method is suitable for more applications with target rotation and size variations. Most importantly, we develop SIS-based method for real-time tracking. The advantage of our tracking algorithm is that we track target without requiring any assumptions to input data, such as Gaussian distribution, motion type and direction, rigid and non-rigid target, and do not need to predict the motion speed and direction (even allow rotation in depth) as our approach continually select good feature for tracking by updating weight at each computation. Another advantage is our tracking is intended for verification – the model templates by Gabor filter can be easily incorporated into the tracker because of its formulation and parameterization. Low computational cost is also one advantage of the algorithm. In our experiments, the tracker processes input data in real-time on an ordinary PC (CPU: Intel P4 - 2GHz, and Memory: 1024 MB).

Acknowledgements

This work was supported in part by a grant from the U. S. Missile Defense Agency, HQ0006-05-C-7277. The authors would like to thank Mr. Richard Felner and Mr. Steve Waugh for their helpful suggestions and comments.

References

1. Biometric Systems Technology, Design and performance Evaluation, Wayman, J.; Jain, A.; Maltoni, D.; Maio, D. (Eds.), Springer, (2005)
2. Handbook of Fingerprint Recognition, Maltoni, D.; Maio, D.; Jain, A.; Prabhakar, S. Springer, (2003)

3. Handbook of Face Recognition, Li, Stan Z, Jain, A.; (Eds.), Springer, 2005.
4. D. A. Clausi and H. Deng: Fusion of Gabor filter and co-occurrence probability features for texture recognition," IEEE Trans. of Image Processing, Vol. 14, No. 7, (2005) 925-936
5. L. Wiskott, J. Fellous, N. Kruger, C. Malsburg: Face recognition by Elastic Bunch Graph Matching, Intelligent Biometric Techniques in Fingerprint and Face Recognition, CRC Press, Chapter 11, (1999) 355-396
6. A. Baumberg and D. Hogg: An efficient method for contour tracking using active shape models, In Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects, (1994) 194-1999
7. M. Black and A. Jepson: Eigen Tracking: Robust matching and tracking of articulated objects using a view-based representation, Int. J. computer Vision, vol. 20, (1998) 63-84
8. D. Freedman and M. Brandstein: A subset approach to contour tracking in clutter, In Proc. Int. Conf. Computer Vision (1999)
9. D. Huttenlocher, J. Noh, and W. Rucklidge: Tracking nonrigid objects in complex scenes, In Proc. Int. Conf. Computer Vision (1993)
10. M. Isard and A. Blake: Contour tracking by stochastic propagation of conditional density, in Proc. Eur. Conf. Computer Vision (1996)
11. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland: Pfinder: Real-time tracking of the human body, IEEE Trans. Pattern Anal. Machine Intell., Vol. 19 (1997) 780-785
12. Y. Yacoob and L. Davis: Tracking rigid motion using a compact-structure constraint, in Proc. Int. Conf. Computer Vision (1999)
13. J. G. Daugman, "Complete discrete 2D Gabor transform by neural networks for image analysis and compression," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 36(7) (1988) 1169-1179
14. R. L. DeValois and K. K. DeValois, Spatial Vision, Oxford Press (1988)
15. J. P. Jones and L. A. palmer: An evaluation of the 2D Gabor filter model of simple receptive fields in cat striate cortex, J. of Neurophysiology, Vol. 58 (1987) 1233-1258
16. A. J. Bell and T. J. Sejnowski: The independent components of natural scenes are edge filters, Vision Research, Vol. 37(23) (1997) 3327-3338.
17. B. A. Olshausen and D. J. Field: Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature, Vol. 381 (1996) 607-609
18. M. Potzach, N. Kruger, and V. D. Malsburg: Improving object recognition by transforming Gabor filter responses, Network: Computation in Neural Systems, Vol. 7(2) (1996) 341-347.
19. S. E. Grigorescu, N. Petkov, and P. Kruizinga: Comparion of texture features based on Gabor filters, IEEE Trans. on Image Processing, Vol. 11(10) (2002)
20. D. Beymer: Face recognition under varying pose, In Proc. IEEE Conf. Computer Vision and Pattern Recognition, (1994) 756-761
21. R. M. Dufour, E. L. Miller, N. P. Galatsanos: Template matching based object recognition with unknown geometric parameters, IEEE Trans. on Image Processing, Vol.11(12) (2002)
22. A. Abu-Naser, N. P. Galatsanos, M. N. Wernick, and D. Schonfeld: Object recognition based on impulse restoration using the expectation maximization algorithm, J. Opt. Soc. Amer. A: Opt. Image Sci., Vol. 15 (1998)
23. J. Liu and R. Chen: Sequential Monte Carlo methods for dynamic systems, J. Am. Stat. Assoc. Vol. 93 (1998) 1031-1041
24. R. Hartley and A. Zisserman: Multiple View Geometry in Computer Vision, Second Edition, Cambridge Press (2003)