Visual Tracking via Sparse Representation Based Linear Subspace Model

JiXiang Zhang*, WeiLi Cai†, Yuan Tian* and YiPing Yang*
*Integrated Information System Research Center
Institute of Automation, Chinese Academy of Sciences
Beijing, P.R. China
{jixiang.zhang,yuan.tian,yiping.yang}@ia.ac.cn
†Logistics Command Academy
Beijing, P.R. China

***Abstract*—The modeling of the object appearance is one of the key issues in the development and application of effective object tracking. This paper presents a tracking algorithm based on representing the appearance of the object using a sparse representation based subspace model. the sparse representation theory offers us a powerful tool to model the object by only a small fraction of the training set. The multi-part subspace appearance model(MSAM) is learned via l^1-minimization and the Gramm-Schmidt process given enough training samples (overcomplete dictionary). Furthermore, a novel model updating strategy is designed to incrementally update the proposed subspace model and the dictionary. Finally, an observation model integrating both sparsity and the likelihood information is designed to embed the proposed modeling approach into the particle filter framework for efficient object tracking. Experimental results demonstrate the robustness and effectiveness of the algorithm, especially when the images are noisy or the objects exhibit large appearance changes.**

***Keywords*-visual tracking; subspace model; sparse representation;**

I. Introduction

Object tracking is a classical problem in computer vision. The purpose is to estimate the trajectory of the object from measured image sequence. It has been extensively studied by researchers and widely applied to visual applications including surveillance, robotics, human-machine interface, intelligent transportation and object based video coding, etc.

The fundamental and challenging task in designing a robust visual tracking algorithm is handling inevitable appearance variations of the tracked object. The variations arise from various factors, e.g., pose variation and shape deformation of the object, illumination change, camera motion, camera viewpoint change, and partial occlusion. Therefore, a flexible model or representation which can adapt to appearance changes plays a critical role in visual tracking.

In the literature, there exists a variety of tracking algorithms attempting to model the appearance variation of the object. Black et al. [1] employ a pre-trained view-based eigenbasis representation and a robust error norm for visual tracking. A subspace constancy assumption is defined for motion estimation instead of relying on the constant brightness assumption. The algorithm can track objects undergoing both changes in viewpoint and changes in pose, but it does not work well under a varying lighting condition. Hager and Belhumeur [2] present a tracking algorithm that is insensitive to illumination variations. A set of illumination basis is constructed at a fixed pose in order to account for appearance variation of the target due to lighting changes. Therefore, it is not clear whether this method works well when both pose and illumination varies. Ho et al. [3] presents a tracking algorithm based on representing the appearances of objects using affine warps of learned linear subspaces of the image space. The tracker adaptively updates this subspace while tracking by finding a linear subspace that best approximates the observations made in the previous frames. Black et al. [4] represent and recover the appearance changes in consecutive frames using a mixture model, and Jepson et al. [5] propose a more elaborate mixture model with an online EM algorithm, however, their altorithm threats the object at pixel level, special care is needed to handle external illumination, such as passing the image through steerable filters that are insensitive to illumination variation.

Although considerable work has already been done above, and significant progress has been achieved, effort in devising tracking algorithm that is applicable in a wide range of conditions is still intensively needed.

Recently, Sparse representation has received a great deal of attentions in the statistical signal processing community[6][7]. It has been successfully applied to image denoising, image compression, feature selection, and so on. In this theory, the signal representation problem is formulated as computing sparse linear representation with respect to an overcomplete dictionary of base elements or signal atoms. Huang and Aviyente [8] present a theoretical framework for signal classification with sparse representation which can deal with signal corruptions such as noise, missing data and outliers. Wright et al.[9] present a sparse representation based face recognition algorithm, they exploited the discriminative nature of sparse representation to perform classification.

Inspired by their work, we develop a tracking algorithm in this paper based on the sparse representation to tackle the problem of appearance variation. The appearance of the

978-0-7695-3836-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CIT.2009.63

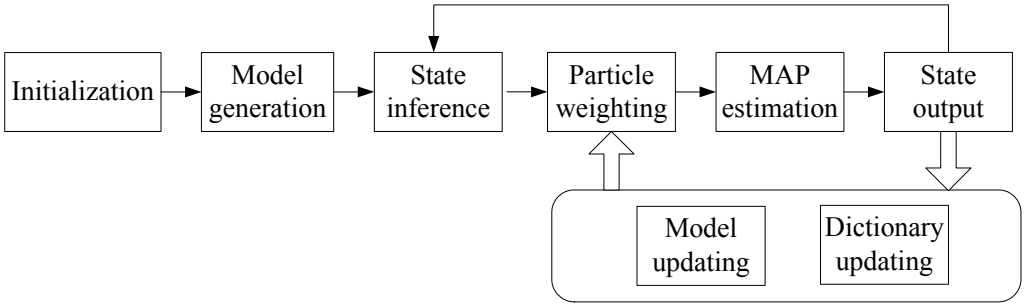

Figure 1. The flow chart of the proposed tracking algorithm

target is modeled using a linear subspace, and the subspace is computed by applying sparse representation over a collection of training images. The appearance of the object is learned online, and updated incrementally over the time. The main contributions of the proposed algorithm are summarized as follows:

- The appearance of the object is modeled by a multi-part subspace model learned using sparse representation by ℓ^1-minimization, and the Gramm-Schmidt process.
- A model updating strategy includes dictionary updating and subspace updating is proposed to enable the subspace to capture the variational appearance changes over time.
- An effective likelihood function based on the learned subspace is developed, and Bayesian inference framework is adopted to improve the accuracy and efficiency of the algorithm.

II. Tracking based on Sparse Representation

A. Overview of the approach

Bayesian inference has been proved to be a flexible and effective tracking framework [10]. Therefore, we embed the MSAM based approach into Bayesian inference framework to build a robust tracking algorithm. The flowchart for our tracking framework using MSAM is illustrated in Fig.1. First, the position of the object is manually initialized, and then the proposed MSAM approach is implemented to model the object. After that, the particle generation process is employed to predict the object state, and each particle is then weighted by the observation model, which is based on the MSAM representation and learned over time. The final object state is obtained by maximum a posterior(MAP) estimate. Finally, the overcomplete dictionary and the object model are incrementally updated according to the final result.

B. Brief introduction of sparse representation

In this section, we give a brief introduction of the sparse representation of a signal in a given overcomplete dictionary[8]. Given a $m \times n$ matrix A, whose columns are the elements of an overcomplete dictionary, and a signal $y \in R^m$, $m < n$. The problem of sparse representation is to find an n dimensional coefficient vector $x \in R^n$, such that $y = Ax$ and the ℓ^0-norm of x is minimized, i.e.,

$$x_0 = \arg\min ||x||_0 \quad \text{subject to} \quad y = Ax, \tag{1}$$

where $|| \cdot ||_0$ denotes the ℓ^0-norm, which is the number of nonzero entries in a vector. However, the problem of finding the sparsest solution of an underdetermined system of linear equations is NP-hard. Recent development of sparse representation shows that, if certain conditions on the sparsity is satisfied, i.e., the solution is sparse enough, The solution of Eq.(1) is equivalent to the solution of the following equation.

$$x_1 = \arg\min ||x||_1 \quad \text{subject to} \quad y = Ax, \tag{2}$$

where $|| \cdot ||_1$ denotes the ℓ^1-norm, which is the sum of absolute values of the vector entries.

Since real data are noisy, the underdetermined system is modified as

$$y = Ax + z, \tag{3}$$

where $z \in R^m$ is a noise term with bounded energy $||z||_2 < \epsilon$, and the sparse solution x can be approximately recovered by solving the following optimization problem:

$$x_1 = \arg\min ||x||_1 \quad \text{subject to} \quad ||y - Ax||_2 < \epsilon, \tag{4}$$

C. Sparse representation based linear subspace learning

In this paper, we formulate the object as the sparse representation in the dictionary, and the dictionary is composed of the training samples. Both object and the training samples are regularized to the same size $w \times h$, and then transformed to a $m = w \times h$ dimensional vector by concatenating the columns.

According to [9], it is reasonable to assume that, if sufficient training samples are available, it will be possible

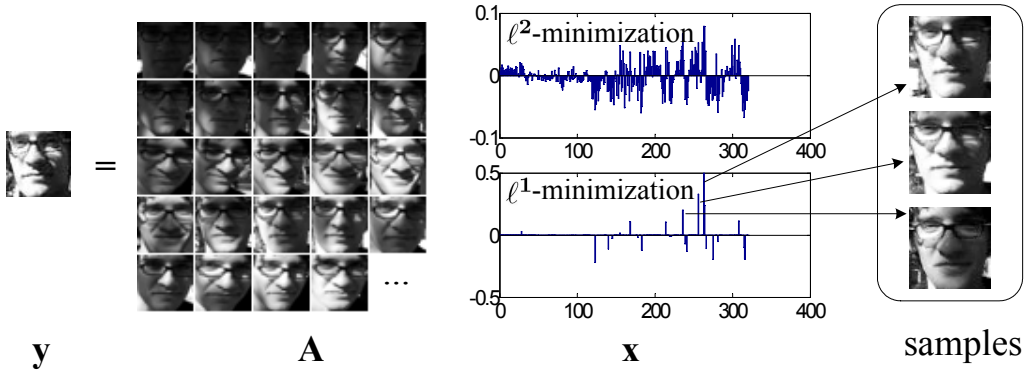

Figure 2. Illustration of the sparse representation: y represents the manually initialized object; A represents the training sample set; x is the coefficients vector via ℓ^1-minimization ℓ^2-minimization. we can see that the ℓ^1-minimization is sparse. the samples related to the three largest coefficient of the sparse solution are shown on the right.

to represent the object as a linear combination of just those training samples that are similar to it. In other words, only a small fraction of the coefficients in x are nonzero, i.e. y can be sparsely represented using A and x. Fig.2 illustrates the sparse representation in object representation. As we can see, most of the entries of the solution x via ℓ^1-minimization are zero or approximately zero, and object y can be represented by the combination of a small number of training samples.

The subspace used to model the object is formed by selecting d samples related to the largest d coefficients. Note that the basis vectors in the dictionary are not orthogonal, in order to compute the distance between the object and the test sample, the Gramm-Schmidt process is applied to orthogonalize the vectors, then the subspace L is achieved.

As the object usually exhibits different appearance properties in different parts, take football player in Fig.3 for example, the appearance of lower part changes more frequently than the upper part due to pose variation. In order to tackle this problem, We divide the object region into $p \times q$ object blocks, each of which is represented by the sparse representation of the training samples related to that region.

Given a sequence of training samples $\mathcal{I} = \{I^t \in R^{W \times H}\}_{t=1,2,\ldots,N}$, we divide the sample I^t into $p \times q$ blocks, as shown in Fig.3, where W and H denote the normalized sample size, $W = w \times q$ and $H = h \times q$, p and q denote block number. In the football player image sequence, the sample size is normalized to 8×24, and divided into 1×3 blocks. i.e. $w = 8, h = 8, W = 8, H = 24, p = 1$ and $q = 3$. For each block $B_{ij} \in R^{w \times h}$, $\mathcal{B}_{ij} = \{B_{ij}^t\}_{t=1,2,\ldots,N}$ are training samples related to block (i,j), and the dictionary A_{ij} is formulated as:

$$A_{ij} = [vec(B_{ij}^1), vec(B_{ij}^2), \ldots, vec(B_{ij}^N)] \tag{5}$$

where $vec(\cdot)$ is an operator transforming a matrix into a $m = w \times h$ dimensional column vector. Finally, the object appearance is formulated as $p \times q$ subspaces $\mathbb{L} = \{L_{ij}, i = 1, \ldots, p.j = 1, \ldots, q\}$, and the subspaces are learned by the sparse solution x_{ij} solved by l^1-minimization.

$$y_{ij} = A_{ij} x_{ij}. \quad i = 1, \ldots, p. \quad j = 1, \ldots, q \tag{6}$$

D. Likelihood evaluation

Once the subspaces $\mathbb{L}$ are achieved, we can use them to calculate the similarity between the object and the candidate sample y'.

The distance between candidate sample part y'_{ij} and the learned subspace L_{ij} is determined by the reconstruction error norm:

$$r_{ij} = ||vec(y'_{ij}) - \overline{L}_{ij} - L_{ij} \cdot L'_{ij} \cdot (vec(y'_{ij}) - \overline{L}_{ij})|| \tag{7}$$

where $\overline{L}_{ij}$ is the column mean of subspace L_{ij}. Thus, the likelihood l_{ij} is computed as

$$l_{ij} \propto \exp(-dis) \tag{8}$$

Here, we define a new vector $\hat{x_{ij}}$ which contains just the largest d coefficients of x_{ij}, and $\hat{A_{ij}}$ contains the corresponding sample vectors.

The sparse representation residual $R_{ij} = ||y_{ij} - \hat{A_{ij}} \hat{x_{ij}}||$ is different, because different parts exhibit different complexity. By reserving the same number of representative samples, the smaller the residual R_{ij}, the more accuracy the sparse representation models the block. we give each part a weight according to the sparse representation residual, indicating the confidence of the subspace L_{ij}.

$$w_{ij} \propto \exp(-R_{ij}) \tag{9}$$

Finally, the overall likelihood between a candidate object part y'_{ij} and the learned subspace model L_{ij} is computed as:

$$W_{ij} \propto l_{ij} \cdot w_{ij} \tag{10}$$

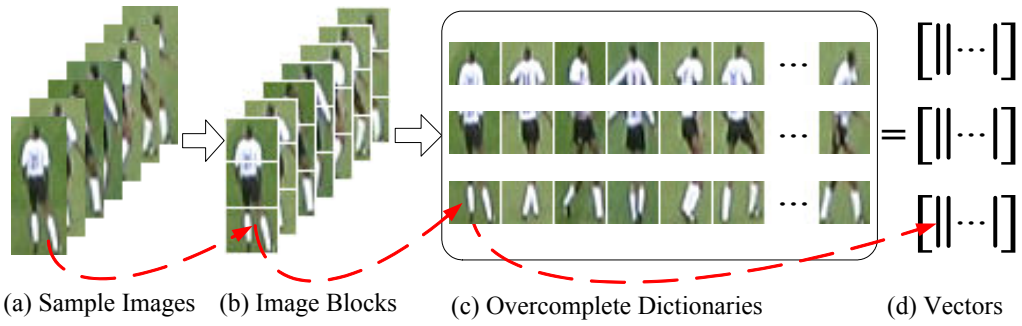

Figure 3. Illustration of the overcomplete dictionary. First, the samples are divided into blocks; Second, the image patches are transformed into vectors; Finally, the dictionary related to each part are obtained.

The overall likelihood between the candidate object and $\mathbb{L}$ is formulated as:

$$W \propto \prod_{1 \le i \le p, 1 \le j \le q} W_{ij} \tag{11}$$

E. MAP Bayesian Tracking

Object tracking can be formulated as a Bayesian inference process for estimating the unknown state X_t. The posterior distribution $p(X_t|Z_t)$ over X_t is recursively updated given all observations $Z^t = \{Z_1, Z_2, \ldots, Z_t\}$ up to time t.

$$p(X_t|Z^t) \propto p(Z_t|X_t) \int p(X_t|X_{t-1}) p(X_{t-1}|Z^{t-1}))dX_{t-1} \tag{12}$$

where $p(X_t|X_{t-1})$ denotes the dynamic model, predicts the object state X_t at time t given the previous state X_{t-1}. $p(Z_t|X_t)$ expresses the observation model, indicates the probability we will observe the measurement Z_t given the state X_t. These two models decide the entire tracking process.

In particle filter, the posterior $p(X_{t-1}|Z^{t-1})$ is approximated as a set of weighted samples $\{X^i_{t-1}, \pi^{(i)}_{t-1}\}^N_{i=1}$, where $\sum_{i=1}^N \pi^{(i)}_{t-1} = 1$, and the Bayesian tracking is represented as:

$$p(X_t|Z^t) \approx kp(Z_t|X_t) \sum_i \pi^{(i)}_{t-1} p(X_t|X^{(i)}_{t-1}) \tag{13}$$

where k is a normalization constant that is independent of X_t. Particle filter can be viewed as an importance sampler for $p(X_t|Z^t)$. First, draw N particles from the proposal distribution

$$X^{(i^+)}_t \sim q(X_t) = \sum_{i^-=1}^{N} \pi^{(i^-)}_{t-1} p(X_t|X^{(i^-)}_{t-1}), \tag{14}$$

and then calculate the particle weight

$$\pi^{(i)}_t = p(Z_t|X^{(i)}_t) \tag{15}$$

In this paper, The object state is represented by a 4 dimensional vector $X_t = [x, y, s, \beta]$, where x, y, s, β denote the x, y translations, the scale, and the aspect ratio. We use a Gaussian distribution to model the state transition distribution as:

$$p(X_t|X_{t-1}) = \mathcal{N}(X_t; X_{t-1}, \Sigma) \tag{16}$$

where Σ is a diagonal covariance matrix whose diagonal elements are σ^2_x, σ^2_y, σ^2_s, σ^2_β. The observation model $p(Z_t|X_t)$ is formulated as $p(Z_t|X_t) \propto W$. where W is defined in Eq.(11), which indicates the likelihood between a candidate sample and the subspace model.

Finally, the state estimate $\hat{X}_t$ can be obtained by maximum a posterior estimation(MAP).

$$\hat{X}_t = \arg\max_{X_t} p(X_t|Z^t) \tag{17}$$

F. Model and dictionary updating strategies

Although it is reasonable to assume that the appearance of the object remains the same between two consecutive frames, the object model would become unsuitable for tracking in scenario where the appearance of the object has been undergoing huge variations due to pose variation and illumination changes. A solution to this problem is online model updating. however, inaccurate updating will introduce the background patch into the model, leading to drift away of the tracking result. Thus, a proper designed updating scheme is crucial for achieving robust tracking performance.

The updating strategy is described as follows: For each part of the tracked object, if its likelihood is higher than λ_1, we directly add it into the dictionary, the object parts around it with the weight higher than λ_2 can also be added into the dictionary, where λ_1, λ_2 are predefined thresholds, and $\lambda_1 > \lambda_2$. The parts that satisfy $\lambda_2 < W_{ij} < \lambda_1$ will be considered to have large appearance changes due to object pose changes or illumination variation, so the sparse representation of these parts have to be recomputed in order to model the object appearance accurately. The strategy will

prevent the appearance changes due to partial occlusion from updated to the model and the dictionary.

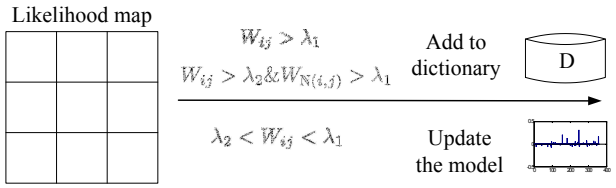

Figure 4. Illustration of dictionary updating. Parts that satisfies $W_{ij} > \lambda_1$ or $W_{ij} > \lambda_2 \& W_{\mathbb{N}(i,j)} > \lambda_1$ are added into the dictionary. Parts that satisfies $\lambda_2 < W_{ij} < \lambda_1$ have to recompute the sparse representation. $W_{\mathbb{N}(i,j)} > \lambda_1$ means that at least one neighbor of part ij has the weight bigger than λ_1.

III. Experimental Results

In this section, we conduct experiments on publicly available databases for object tracking, for the purpose of demonstrating the efficiency of the proposed algorithm. In the experiment, we first generate the training set from the a portion of images in the sequence, and test the algorithm using the rest of the images. The object is initialized manually in the first frame, the particle number is set to 400, and the covariance matrix Σ is assgined as $\text{diag}(4^2, 4^2, 0.02^2, 0.002^2)$,

The first experiment is to track a outdoor football player from PETS2003 dataset. The appearance of the target changes over time due to pose variation, especially the lower part. In the experiment, $w = 8, h = 8, p = 1$ and $q = 3$ The tracking results are shown in Fig.5.

The second experiment is to test the algorithm under low quality image sequence contaminated by noise. The image sequence used in this experiment is obtained by adding Gaussian random noise to images used in the first experiment. As shown in Fig.6, The results demonstrate that the algorithm is robust even when the data is very noisy.

The third experiment is to test the performance of the algorithm when the object appearance has abrupt changes because of illumination changes. In this experiment, $w = 6, h = 6, p = 3$ and $q = 3$, The length of the image sequence is 501. Frame 1 to 320 are used for generating the dictionary. The tracking result is shown in Fig.7.

The experimental results show that the proposed algorithm can adapt to appearance changes flexibly. Here, a brief explanation is offered. Consider Fig.8, each circle represents a sample in the dictionary. The samples in the dictionary form many subspaces according to different appearances. Given an object part, the subspace it belongs to is learned by the proposed method. As the appearance changes, the likelihood becomes small, and the model updating is invoked to find the new subspace. Although the discussion above is rather informal, The geometric explanations do provides some validity for the robustness of the proposed algorithm against object appearance changes arising from various factors.

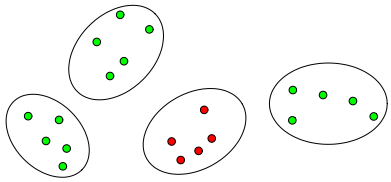

Figure 8. Geometric explanation of the proposed algorithm. each circle denotes a sample in the dictionary, and the subspace model is obtained by the proposed algorithm, denoted as red circles. different subspace related to different appearances

IV. Conclusion

In this article we have introduced an approach for object tracking in image sequence using a multi-part subspace based representation. In our approach, the subspace of object parts are learned by sparse representation through l^1-minimization, and the Gramm-Schmidt process. Both the dictionary and the object model are updated online adaptively. To tracking the object in the particle filter framework, we develop the observation model based on the similarity between the candidate object the the subspace model. The proposed algorithm is insensitive to noise, illumination changes and pose variation. Experimental results have demonstrated the efficiency and effectiveness of the proposed algorithm.

References

[1] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[2] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 403, 1996.

[3] J. Ho, K. C. Lee, M. H. Yang, and D. Kriegman. Visual tracking using learned linear subspaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 782–789, 2004.

[4] Michael J. Black, David J. Fleet, and Yaser Yacoob. A framework for modeling appearance change in image sequences. In *International Conference on Computer Vision*, pages 660–667, 1998.

[5] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311, 2003.

[6] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l(1)-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

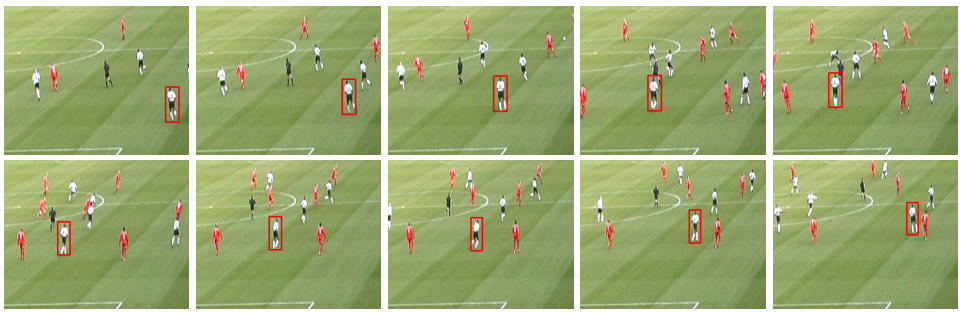

Figure 5. Tracking results of football player image sequence

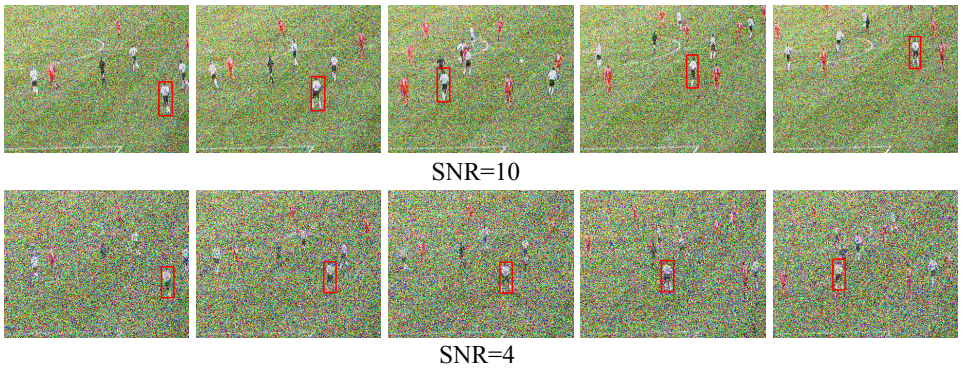

Figure 6. Tracking results with different signal-to-noise ratio

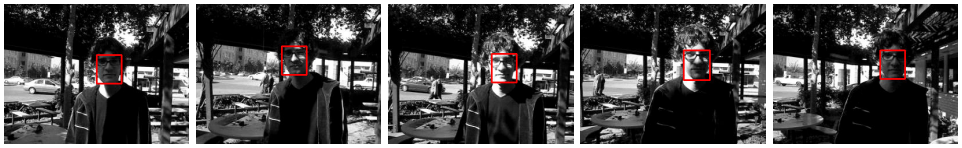

Figure 7. Appearance changes due to illumination variation

[7] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[8] Ke Huang and Selin Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems*, pages 609–616. 2007.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[10] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.