Computer Vision and Image Understanding 113 (2009) 345–352

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# Object tracking using SIFT features and mean shift

Huiyu Zhou [a,*], Yuan Yuan [b], Chunmei Shi [c]

[a] Brunel University, Uxbridge, Electronic and Computer Engineering, Middlesex UB8 3PH, UK
[b] Aston University, Aston Triangle, Birmingham B4 7ET, UK
[c] People's Hospital of Guangxi, Nanning, Guangxi, China

ARTICLE INFO

*Article history:*
Received 2 November 2007
Accepted 18 August 2008
Available online 29 August 2008

*Keywords:*
Object tracking
Color histogram
Mean shift
SIFT features
Expectation–maximization

ABSTRACT

A scale invariant feature transform (SIFT) based mean shift algorithm is presented for object tracking in real scenarios. SIFT features are used to correspond the region of interests across frames. Meanwhile, mean shift is applied to conduct similarity search via color histograms. The probability distributions from these two measurements are evaluated in an expectation–maximization scheme so as to achieve maximum likelihood estimation of similar regions. This mutual support mechanism can lead to consistent tracking performance if one of the two measurements becomes unstable. Experimental work demonstrates that the proposed mean shift/SIFT strategy improves the tracking performance of the classical mean shift and SIFT tracking algorithms in complicated real scenarios.



## 1. Introduction

Visual object tracking is an important topic in multimedia technologies, particularly in applications such as teleconferencing, surveillance and human–computer interface. The goal of object tracking is to determine the position of the object in images continuously and reliably against dynamic scenes [1]. To achieve this target, a number of elegant algorithms have been established. For example, considering Gaussian and linear problems, Welch and Bishop [2] presented a Kalman filter-based method for tracking a user's pose for interactive computer graphical. The proposed single-constraint-at-a-time (SCAAT) tracking utilized single observations from optical sensors and fused the measurements from different sensors in order to improve the tracking accuracy and stability. As a promising solution to non-Gaussian and non-linear systems, particle filter-based approaches have been included in current tracking technologies. These schemes, e.g., [3,4], recruited particles for computing a sampled representation of the posterior probability distribution over scene properties of interest, based on image observations. Other tracking strategies can also be found as Multiple Hypothesis Tracking (e.g., [5,6]), kernel-based tracking (e.g., [7,8]), and optical flow-based tracking (e.g., [9]).

In spite of successes in many real circumstances, these established algorithms face challenges from severe image occlusions and background clutters, where sometimes the trackers cannot effectively converge to the real settlement. Recently, mean shift was proposed as one of the efficient tools to handle partial occlusions and significant clutters [10,11]. The mean shift algorithm was designed so as to search for a local probability density function (PDF) that approximates the empirical PDF. Mean shift has shown promising performance in many circumstances against image occlusions and clutters. In spite of its success in specific applications, however, mean shift has presented less efficiency in the presence of dramatic intensity or color changes in a pre-defined neighborhood [12]. In these situations, additional features may be used as a complement that can improve the capability of the trackers. Unfortunately, edges, corners and silhouette are application-depended features, and cannot effectively work in the case of variable scaled, rotated or translated images. In recent years, some of the decent techniques have been used to seek optimal solutions to these problems. For example, scale invariant feature transform (SIFT) [13] was used to generate feature points in a full image. These feature points appear invariant to any scaling, rotation or translation of the images. Therefore, the SIFT features can be integrated with an established tracking system for improving the performance of the latter.

In this paper the proposed tracking algorithm is an effective integration of mean shift and SIFT feature tracking. The proposed approach will apply a similarity measurement between two neighboring frames in terms of color and SIFT correspondence. Technically, a track will be made if mean shift and SIFT feature tracking lead to approximate probability distributions (e.g., intensity and color) within the corresponding region in the next image frame (ideally the two probability distributions should be identical if the scenario does not change). An expectation–maximization algorithm is employed in order to pursue a maximum likelihood estimate using the measurements from mean shift and SIFT

* Corresponding author. Fax: +44 1895 232806.
*E-mail address:* Huiyu.Zhou@brunel.ac.uk (H. Zhou).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.08.006

correspondence. The main contributions of this paper consist of: (1) a combinatorial theory of mean shift and SIFT feature matching is proposed for object tracking, and (2) the tracking performance of the proposed strategy can be experimentally justified against the classical algorithms, i.e., mean shift and SIFT tracking.

The rest of the paper is organized as follows: in the next section, literature review is conducted for outlining current research progress. In Section 3 a Lagrangian-based object function is introduced in order to ideally adapt the ellipse to find the real location of the object. Experimental work is conducted in Section 4, and finally conclusions are given in Section 5.

## 2. Literature review

Object tracking, according to its properties, can be categorized into three groups: feature-, model-, optical flow-based approaches. These three classes, to some degree, can find common application areas. Their individual characteristics will be summarized in the following sections, where algorithmic success and features are briefly introduced (comprehensive review can be found in relevant documentations). First, feature-based approaches are introduced.

### 2.1. Feature-based

Feature-based algorithms were originally developed for tracking a small number of salient features in a long sequence. They involve the extraction of regions of interest (features) in the images and then identification of the counterparts in individual images of the sequence. Obviously, this strategy seeks very specific correspondences across frames and hence may reduce mis-tracks due to the unique characteristics of the features (this may not be true in presence of multiple similarities). Typical feature-based tracking algorithms are: multiple hypothesis tracking (MHT), hidden Markov models (HMM), artificial neural network (ANN), particle filter, Kalman filtering (KF) and mean shift (MF).

The MHT algorithm was originally presented by Reid [5]. The concept behind the algorithm is an iterative process that forms a feedback to "tune" the match between two feature groups until the match arrives at an optimal agreement up to a particular criterion. This algorithm considers multiple tracking candidates and intends to find the best fit to the real image descriptors. It works well in multiple object tracking. However, the classical MHT technique by itself is computationally expensive both in time and memory [6], which reluctantly supports real applications.

A HMM is normally used to extract the transformation between two images or moving 3D structures in object tracking. However, this is not deterministic, and since the model is hidden, there may in fact be more than one possibility of transformation that results in the feature positions. Thus the most likely sequence of transitions is sought. Using algorithms such as the Baum–Welch algorithm and its modifications [14], people train HMM by adjusting the weights of the transitions to better model the relationship of the actual training samples. HMM-based approaches do not require analytical solutions to certain problems, being effective in handling very complicated environments. Nevertheless, the required training stage in HMM must be supervised and it is difficult to apply a pre-trained HMM for the overall applications. Similarly, a ANN also needs to determine its weights by training, although ANN methodology has been optimistically applied to object (or motion) tracking [15,16].

Comparing particle filter to Kalman filter, the former has a more robust performance in the case of non-Gaussianity and non-linearity due to the simulated posterior distribution. In a particle filter, a large number of particles are desirable to represent the posterior distribution, especially in the situations where new measurements appear in the tail of the prior, or if the likelihood is strongly peaked. To solve the computational problem raised by the large particle numbers, for example, MacCormick and Isard [17] developed partitioned sampling, which requires that the state-space can be sliced. Sullivan et al. [18] proposed layered sampling using multiscale processing of images. It turns out that these solutions significantly reduce the computational costs but in-depth efforts are desirable for better efficiency.

In mean shift tracking algorithms, a color histogram is used to describe the target region. The Kullback–Leibler divergence, Bhattacharyya coefficient and other information-theoretic similarity measures are commonly employed to measure the similarity between the template (or model) region and the current target region. Tracking is accomplished by iteratively finding the local minimum of the distance measure functions. For example, Yang et al. [19] proposed a simple-to-compute and more discriminative similarity measure in spatial-feature spaces. The new similarity measure allows the mean shift algorithm to track more general motion models in an integrated way. To reduce computational complexity and make a model of linear order they employed the recently proposed improved fast Gauss transform. The state-of-the-art of mean shift has been popularly applied to practical problems, e.g., [20–22]. Due to the efficiency, robustness and stability in tracking of colored objects, mean shift is here adopted in the proposed approach.

### 2.2. Model-based

Strictly speaking, model-based tracking is an example of feature-based tracking. The reason why it is independently described is due to the requirement of grouping, reasoning and rendering, which may defer from the feature-based tracking. In addition, prior knowledge about the investigated models is normally required. For example, in the case of multiple objects tracking the binary representation (models) of the targets must be obtained a-priori. This may be followed by applying a stage of model recognition.

Lowe [23] used the Marr–Hildreth edge detector to extract edges from an image, which were then chained together to form lines. These lines were then matched and fitted to those in the model. The Hough transform was utilized to achieve a similar idea in [24]. Model-based tracking schemes share the same challenges as the feature-based trackers do. For example, occlusion is a significant cause of instabilities, resulting in poor tracking performance. The RAPID tracker [25] handled this situation by use of a pre-computed table of indexed visible features. Any occluded feature will be reported due to its absence from the table. To track human motion in activities such as walking, running, and jumping, Gavrila and Davis [26] matched edges in the image with those of an appearance model using distance transforms. A decomposition approach and a best-first technique were used to search through the high dimensional pose parameter space. A robust variant of Chamfer matching was used as a fast similarity measure between synthesized and real edge images.

### 2.3. Optical flow-based

Optical flow is the vector field which describes how the image changes with time. The two dimensional projection from the three-dimensional velocity field observed by the camera needs to be estimated. However, this has been proved to be extremely difficult to achieve due to problems such as the "aperture effect". To find the optical flow in the image sequence, people attempt to use feature-based, gradient-based, or correlation-based approaches. Most these approaches are of intensive computation, and hence computational optimization is demanding.

Optical flow methods are normally used for generating dense flow fields by computing the flow vector of each pixel under the brightness constancy constraint [27]. This computation is often carried out in the neighborhood of the pixel either algebraically [9] or geometrically [28]. Extending optical flow methods to compute the translation of a rectangular region is somewhat achievable. One of the example was reported in [29], where Shi and Tomasi proposed the well-known Shi–Tomasi–Kanade (STK) tracker which iteratively computes the translation of a region centered on an interest point. Once the new location of the interest point is obtained, the STK tracker evaluates the quality of the tracked patch by computing the affine transformation. This scheme works effectively and fast in most circumstances. Further work is a need of reducing incorrect point correspondence.

## 3. Similarity search

### 3.1. Similarity measure by mean shift

Object tracking is equivalent to similarity search across two neighboring image frames. Given the predicted target's position in the current frame and its uncertainty, the measurement task assumes the search of a confidence region for the target candidate that is the most similar to the target model. The similarity measure conducted here is based on color information. The sample points in the current frame are denoted by $\mathbf{I}_x = (\mathbf{x}_i, \mathbf{u}_i)_{i=1}^N$, where $\mathbf{x}_i$ is the 2D coordinates and $\mathbf{u}_i$ is the corresponding feature vector (e.g., RGB colors of sample points). The sample points in the target image are $\mathbf{I}_y = (\mathbf{y}_j, \mathbf{v}_j)_{j=1}^M$, encoding the 2D coordinates and the corresponding feature vector. The targets can be in the joint feature-spatial space.

Given the sample points and the kernel function $k(x)$, the probability density function (p.d.f.) of the object in the current image can be estimated using kernel density estimation [30] as

$$\tilde{p}_x(\mathbf{x},\mathbf{u}) = \frac{1}{N}\sum_{i=1}^{N} w\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right) k\left(\left|\frac{\mathbf{u}-\mathbf{u}_i}{h}\right|^2\right), \tag{1}$$

where $w$ is a weight function, $k$ is a kernel function, $\sigma$ and $h$ are the bandwidths in the spatial and feature spaces, respectively. Similarly, estimation of p.d.f of the target image can be available. Using feature-spatial space helps build up correspondence between the feature and its counterparts.

Similarity measure needs to be achieved for the correspondence between two groups of the sample points. People normally apply the Kullback–Leibler divergence [31] and the Bhattacharyya distance [7] to measure the affinity between two distributions. The former is expressed as

$$\int p_y(u)\log\frac{p_y}{p_x}\mathrm{d}u. \tag{2}$$

The latter is represented as

$$\begin{cases} B(I_x, I_y) = \sqrt{1-\rho(p_x,p_y)}, \\ \rho(p_x,p_y) = \int \sqrt{\tilde{p}_x(u)\tilde{p}_y(u)}\mathrm{d}u. \end{cases} \tag{3}$$

These measures demand two sequential $\mathcal{O}(N^2)$ operations for the p.d.f and integral calculation on the sample points.

To find a mode of $\tilde{p}_x(\mathbf{x},\mathbf{u})$, the gradient of Eq. 1 with respect to the vector $\mathbf{y}$ is derived and then set to be zero:

$$\frac{\partial\tilde{p}_x(\mathbf{x},\mathbf{u})}{\partial\mathbf{u}} = \frac{2}{N}\sum_{i=1}^{N} w\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right) k'\left(\left|\frac{\mathbf{u}-\mathbf{u}_i}{h}\right|^2\right)\Sigma_i^{-1}(\mathbf{u}-\mathbf{u}_i) = 0, \tag{4}$$

where $k' = \mathrm{d}k/\mathrm{d}t$, and $\Sigma_i$ is covariance matrix. In generic situations, the Hessian of $\tilde{p}_x(\mathbf{x},\mathbf{u})$ is

$$\nabla^2\tilde{p}_x(\mathbf{x},\mathbf{u}) = -2c \times \sum_{i=1}^{N} w\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right)\left(k\left(\left|\frac{\mathbf{u}-\mathbf{u}_i}{h}\right|^2\right)I + 2k'\left(\left|\frac{\mathbf{u}-\mathbf{u}_i}{h}\right|^2\right)\right), \tag{5}$$

where $c$ is a constant and $I$ is identity matrix. If $k$ is piecewise constant, the Hessian of $\tilde{p}_x(\mathbf{x},\mathbf{u})$ is

$$\nabla^2\tilde{p}_x(\mathbf{x},\mathbf{u}) = -2c\sum_{i=1}^{N} w\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right) k\left(\left|\frac{\mathbf{u}-\mathbf{u}_i}{h}\right|^2\right) I \tag{6}$$

To solve for $\mathbf{u}$ one needs to employ a fixed iteration scheme as $\mathbf{u}^{m+1} = f(\mathbf{u}^m)$ ($m$ is index), assuming isotropic covariances, $\Sigma_i = \sigma^2\mathbf{I}$

$$f(\mathbf{u}) = \sum_{i=1}^{N} \frac{k'\left(\left|\frac{\mathbf{u}-\mathbf{u}_i}{h}\right|^2\right)}{\sum_{i'=1}^{N} k'\left(\left|\frac{\mathbf{u}-\mathbf{u}_{i'}}{h}\right|^2\right)}\mathbf{u}_i, \tag{7}$$

where the vector $f(\mathbf{u}) - \mathbf{u}$ is the mean shift.

Now, let us take a look at the convergence properties of mean shift. Cheng [20] claimed that mean shift is an instance of gradient ascent with an adaptive step size. He discovered that unlike Newton's method, each iteration of mean shift is guaranteed to bring us closer to a stationary point. On the other hand, like Newton's method, mean shift can get stuck at a saddle point or mistake a start at a local minimum for a local maximum [22]. Most likely, a number of discontinuities can happen when $k'$ of Eq. 7 is sought [22]. One of the method to avoid these discontinuities is to take infinitesimal steps for moving the direction of the local gradient. Unfortunately, if the step size is too large, the rate of convergence cannot be guaranteed [21]. In this paper, the employment of SIFT feature correspondence may lead to an optimal solution to this problem.

### 3.2. SIFT feature corresponding

For discrete data, appropriate selection of the employed kernel can determine the number of steps [21]. In other words, the convergence depends on the characteristics of $k'$. Evidence shows that if $k'$ is the uniform kernel, then convergence may just take a finite number of iterations before an optimal correspondence is reached a good location. As a result, one can assume that in case $k'$ tends to be non-uniform, additional terms can be combined with $k'$ so as to make the final kernel to be pseudo-uniform. SIFT feature correspondence is such a scheme that brings a component to the formulation of the final $k'$. Before proceeding this new strategy, let us firstly summarize the basics of how SIFT works.

#### 3.2.1. SIFT theory

SIFT attempts to extract scale-invariant features by using a staged filtering approach [13]. The first stage identifies key locations in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. The features hold partial invariance to local variations, such as affine or 3D projections, by blurring image gradient locations. The resulting feature vectors are called SIFT keys. The feature extraction can be computed very efficiently by building an image pyramid with re-sampling between each level.

Let the scale space of an image be $L(x,y,\sigma_L)$, resulting from the convolution of a variable-scale Gaussian, $G(x,y,\sigma_L)$ for an image $I(x,y)$:

$$L(x,y,\sigma_L) = G(x,y,\sigma_L)\bigstar I(x,y), \tag{8}$$

where $\bigstar$ is a convolution operation and

$$G(x,y,\sigma_L)=\frac{1}{2\pi\sigma_L^2}\exp^{-(x^2+y^2)/2\sigma_L^2}. \tag{9}$$

Stable key-point locations in scale space can be computed from the difference of Gaussians separated by a constant multiplicative scalar $s$:

$$G(x,y,\sigma_L)=L(x,y,s\sigma_L)-L(x,y,\sigma_L), \tag{10}$$

The difference-of-Gaussian function can be treated as an approximation to the scale-normalized Laplacian of Gaussian $\sigma_L^2\triangledown^2G$ [32]. $D$ can be linked to $\sigma_L^2\triangledown^2G$ by the following equation:

$$\frac{\partial G}{\partial\sigma_L}=\sigma_L\triangledown^2G. \tag{11}$$

Therefore,

$$G(x,y,s\sigma_L)-G(x,y,\sigma_L)\approx(s-1)\sigma_L^2\triangledown^2G. \tag{12}$$

This indicates that the difference-of-Gaussian function has scales differing by a constant factor, while it incorporates the $\sigma_L^2$ scale normalization required for the scale-invariant Laplacian.

### 3.3. SIFT and mean shift-based similarity measure

Referring to Eq. 1, a similarity measure by SIFT is integrated. The p.d.f of the object in the current image is composed of a classical term and the new term from the SIFT contribution:

$$\hat{p}_x(\mathbf{x},\mathbf{u})=\frac{1}{N}\sum_{i=1}^{N}\left(w_1\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right)k\left(\left|\frac{\mathbf{u}-\mathbf{u}_i}{h}\right|^2\right)+w_2\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right)f_s(\mathbf{x},\mathbf{u})\right), \tag{13}$$

where $f_s$ is a Gaussian distribution based on the SIFT feature correspondence, $w_1$ and $w_2$ are two weight functions. These two weights can be estimated on the basis of the Euclidean distance between the real correspondences and the predicted positions. They are updated on pair-wise frames. Therefore, one can also have

$$\sum_{i=1}^{N}\left(w_1\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right)+w_2\left(\left|\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}\right|^2\right)\right)=1. \tag{14}$$

and

$$f_s(\mathbf{x},\mathbf{u})=\frac{1}{2\pi\sigma_s^2}\exp^{-((v_{xi}-v_{x0})^2+(v_{yi}-v_{y0})^2)/2\sigma_s^2}, \tag{15}$$

where $(v_{xi},v_{yi})$ denote the two-dimensional Euclidean distance between the detected region centre using mean shift and the detected feature position via SIFT, $(v_{x0},v_{y0})$ stand for the mean distance along two axes and $\sigma_s$ is variance of the distance.

The weight and kernel functions ($w$ and $k$) can be seriously affected in presence of noise or corresponding errors. Consequently, successful determination of $w_1$ and $w_2$ cannot be guaranteed due to the biased $w$ and $k$ function values. To effectively solve this problem, first, historic $w$ and $k$ data is used for parameter regression using least squares algorithms, e.g., least median of squares [33]. Second, referring to the time-line, we can retain the corresponding $w_1$ and $w_2$ values in current frame from the curve regression.

Considering the object function as shown in Eq. 13, to estimate the mean shift the following E and M stages are iterated using the established expectation–maximization (EM) algorithm [34]. The EM algorithm has been proved to be an effective maximum-likelihood algorithm for learning a Gaussian mixture model. An expectation (E) step consists of evaluating the posterior probabilities for each mixture component. A maximization (M) step then updates the mixture components [35].

(1) E-stage: for isotropic kernels, for example, the image position can be held as:

$$\begin{cases} f(\mathbf{u})=\sum_{r=1}^{N}q(r|\mathbf{u})\mathbf{u}_r, \\ q(r|\mathbf{u})=\frac{p(r|\mathbf{u})\sigma_r^{-2}}{\sum_{r'=1}^{N}p(r'|\mathbf{u})\sigma_{r'}^{-2}}, \end{cases} \tag{16}$$

where $q(r|\mathbf{u})$ is the posterior probability or responsibility $p(r|\mathbf{u})$ re-weighted by the inverse variance and re-normalized. Assuming i.i.d. image points exist, one then has the log-likelihood of image data as

$$\sum_{i=1}^{N}\mathscr{L}=\sum_{i=1}^{N}\log q(\mathbf{u},\mathbf{z}|\eta), \tag{17}$$

where $z$ is a hidden variable, whose value has presumably been known, $\eta$ is a parameter vector, and the expectation with respect to the posterior distribution is

$$Q(\mathbf{z}^{\tau}|\mathbf{z}^{\tau-1})=\sum_{i=1}^{N}\sum_{j=1}^{M}q(\mathbf{z}|\mathbf{u},\eta^{\tau})\log q(\mathbf{u}|\mathbf{z},\eta)+C, \tag{18}$$

where the term $C$ is independent of $\mathbf{z}$.

(2) M-stage: the new estimates are deductive if a maximization is reached

$$\mathbf{z}^{\tau+1}=\arg\max Q(\mathbf{z}|\mathbf{z}^{\tau-1}). \tag{19}$$

To arrive at this maximization, a derivative of $Q$ is taken in terms of $\mathbf{z}$ and then set to be zero:

$$\frac{\partial Q}{\partial\eta}=\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{q(\mathbf{z}|\mathbf{u},\eta^{\tau})}{q(u|\mathbf{z},\eta)}\frac{\partial q(\mathbf{u}|\mathbf{z},\eta)}{\partial\eta}=0. \tag{20}$$

Let $\mathbf{u}_z$ be a mean value, then

$$\frac{\partial q(\mathbf{u}|\mathbf{z},\eta)}{\partial\eta}=q(\mathbf{u}|\mathbf{z},\eta)\Sigma_z(\mathbf{u}-\mathbf{u}_z-\eta). \tag{21}$$

Finally, the solution for $\eta$ is

$$\eta^{\tau+1}=\left(\sum_{i=1}^{n}\sum_{j=1}^{M}q(\mathbf{z}|\mathbf{u},\eta^{\tau})\Sigma_z^{-1}\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M}q(\mathbf{z}|\mathbf{u},\eta^{\tau})\Sigma_z^{-1}(\mathbf{u}_z-\mathbf{u}). \tag{22}$$

### 3.4. Proposed algorithm

The entire algorithmic flowchart can be summarized as follows:

(1) Define a rectangle on the region of interest in the first frame of a video sequence.
(2) Compute the color histogram of this region, whilst extracting SIFT features within this region using Eqs. (1)–(3), (7), (12).
(3) In the second frame, start from the former location and examine the surroundings for similarity measure using Eqs. 2, 3 and 13. The sum of squared difference (SSD) method is applied for SIFT feature correspondence across frames.
(4) Launch the proposed EM algorithm to search for an appropriate similarity region whilst minimizing the distance between the detected locations by mean shift and SIFT correspondence, respectively.
(5) Iterate the above steps till the difference between two mean shifts is smaller than a threshold (i.e., 0.01).

In the implementation, the probability distribution of the object to be tracked is continuously evaluated. Computational instability

**Fig. 1.** Test sequences used in current evaluation.

may be raised due to lost color histograms or SIFT features (e.g., occlusions). In this case, the estimated probability distribution in the previous frame can be assigned more weights and be used to dominate locating the object till the object appears again.

## 4. Experimental work

Four publicly available test sequences [1] are used to evaluate the proposed tracking algorithms (see Fig. 1). These sequences consist of indoors and outdoors testing environments so that the proposed scheme can be fully evaluated. For comparison purposes, classical mean shift tracking [36] and SIFT-based sum of squares (SSD) correspondence approaches [13] are utilized. It must be pointed out that in this evaluation there is no intension to track multiple objects. On the contrary, a single object is detected in the first frame of each sequence, followed by continuous tracking to the remaining part of the sequence. In some sequences, there are more than one object in the scene. These scenarios are set up for evaluating the performance of a tracking system against interference in this "multiple candidate" circumstance.

First of all, sequence "single person in darkness" is tested, in which the person conducts casual movements. This sequence was utilized to evaluate the performance of the proposed tracker in a poor lighting environment. Secondly, to test the proposed algorithm in complicated scenarios we here employ a "four person" sequence, where an object's tracking will be distracted by its neighbors. These two sequences just mentioned are related to the contents including indoors human actions. As an extension of the evaluation, the performance of the proposed tracking scheme is also evaluated in outdoors environments. The third sequence namely "traffic condition" is investigated. In this test, it is interesting to explore the characteristics of the proposed approach to the change of illumination and moving objects. Finally, the proposed SIFT-mean shift tracker is applied to test the "fast movement" sequence, where human objects quickly change their posture. Table 1 illustrates the configuration of each image sequence.

Exemplar tracking results of sequence "single person in darkness" are illustrated in Fig. 2. The first row represents the outcomes of the classical mean shift tracker. Clearly, this tracker led to drifts in such a poor lighting situation. This is due to the fact that the background's color is approximate to that of the human face, which deviates the track. In the second row, the detected SIFT features correctly settle on the face. Hence, the proposed SIFT-mean shift tracker can use these good features for probability distribution computation (see Eq. (13)). As a consequence, the third row demonstrate that the proposed tracker has better performance than the classical mean shift.

**Table 1**
Details of four image sequences used in the evaluation (fps, frames per second)

| Sequences | Size | Frame-number | fps | Object-number |
|---|---|---|---|---|
| 'Single person in darkness' | $720 \times 576$ | 680 | 25 | 1 |
| 'Four person' | $720 \times 576$ | 1006 | 25 | 4 |
| 'Traffic condition' | $720 \times 576$ | 3748 | 25 | >8 |
| 'Fast movement' | $720 \times 576$ | 448 | 25 | 1 |

Fig. 3 shows three image examples of performance comparison of different approaches in sequence "four person". The challenge of this sequence is that a tracking system needs to effectively handle the situation where a male adult was occluded by the others when they crossed over. In this particular example, the aim is to locate the person wearing a white T-shirt, who was holding his arms on the chest. In addition, this person slowly changed his posture during the tracking. This led to gradual disappearing of the feature points. Thereby, the tracking results of mean shift, shown on rows 1 and 2, are incorrectly presented. On the contrary, the proposed SIFT-mean shift algorithm allowed the subject to be optimally tracked (see row 3).

Fig. 4 shows some examples of performance comparison of different approaches in sequence "four person" when object occlusions happen. This figure shows three columns that refer to frames 40, 67 and 76. It is observed that SIFT fails to detect the object with white T-shirt, and mean shift is reluctant to capture the same object in occlusions. Comparably, the proposed SIFT-mean shift properly captures the object throughout the occlusion procedure. In this example, the success of the proposed algorithm attributes to correct identification of the object before the occlusion.

[1] http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html.

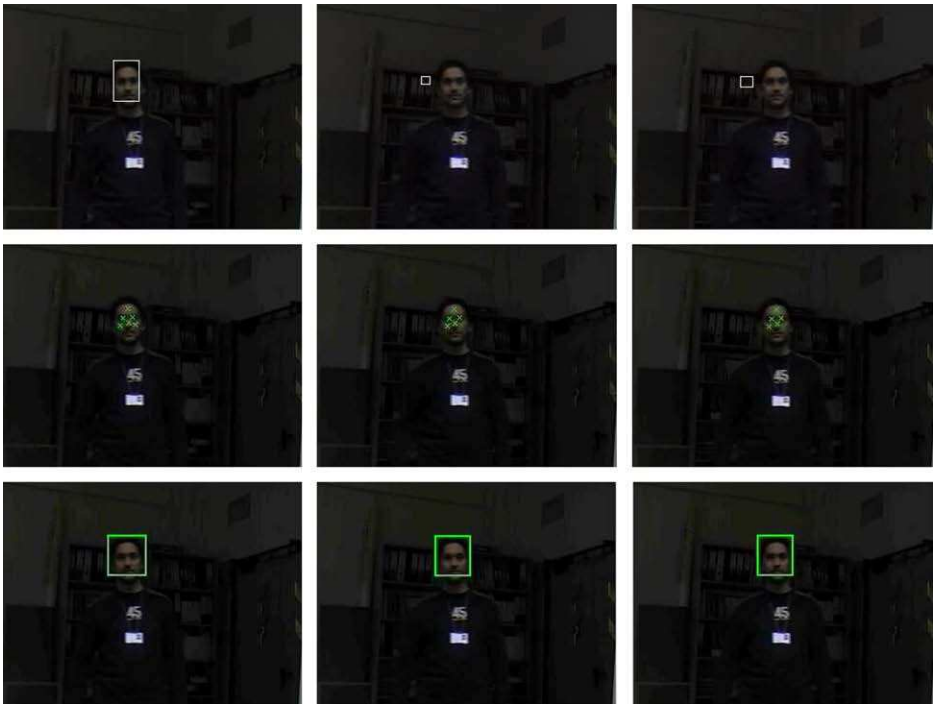

**Fig. 2.** Sequence 1: tracking comparison of the classical mean shift (first row), SIFT feature correspondence (2nd row, SIFT features marked as "×") and proposed tracker (3rd row).

**Fig. 3.** Sequence 2: tracking comparison of the classical mean shift (first row), SIFT feature correspondence (2nd row, SIFT features marked as "×") and proposed tracker (3rd row).

Statistics of the tracking errors in different schemes are here provided (see Table 2). These errors refer to the Euclidean distance between the object detection (centers of bounding boxes are considered) by individual algorithms and the ground truth created by a trained professional. Significantly, the proposed SIFT-mean shift scheme has the lowest error values in different scenarios. Table 2 clearly denotes the tracking performance of individual techniques against the entire test time. In the meanwhile, it is necessary to investigate their temporal performance. For example, Figs. 5, 6 illustrate the performance comparison of

**Fig. 4.** Performance comparison of classical mean shift (first row), SIFT feature correspondence (2nd row, SIFT features marked as "×") and proposed tracker (3rd row) in case the SIFT approach fails in object occlusions.

**Table 2**
Statistics of tracking errors in different scenarios by individual approaches (units: pixels)

| Sequences | Mean shift | SIFT-SSD | SIFT-mean shift |
|---|---|---|---|
| 'Single person in darkness' | 8.2 | 6.9 | 4.6 |
| 'Four person' | 6.8 | 6.1 | 5.2 |
| 'Traffic condition' | 5.1 | 4.4 | 3.6 |
| 'Fast movement' | 2.3 | 2.0 | 1.7 |

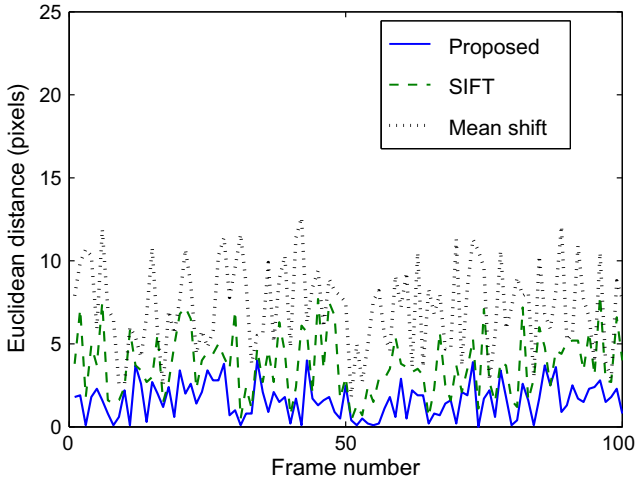


**Fig. 5.** Illustration of tracking accuracy in sequence "single person in darkness": the Euclidean distance between the estimated objection position and the ground truth is plotted against frame numbers.

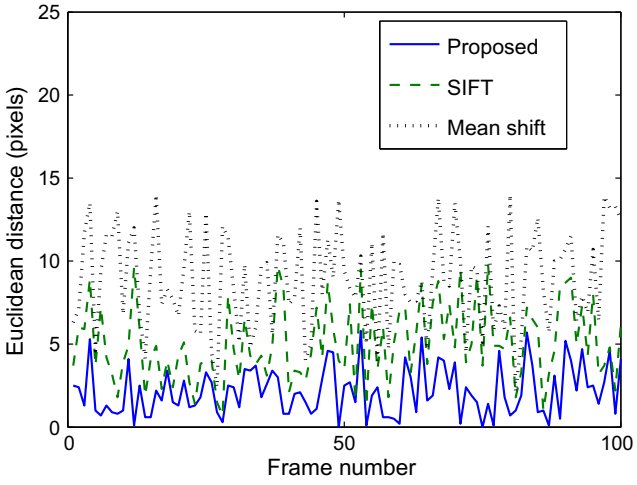


**Fig. 6.** Illustration of tracking accuracy in sequence "traffic condition": the Euclidean distance between the estimated objection position and the ground truth is plotted against frame numbers.

these three approaches in two different sequences, where the proposed SIFT-mean shift possesses less Euclidean errors than others. Note that these sample periods come from intermediate segments of the image sequences.

Finally, with regards to the process speed, it is verified that three algorithms, mean shift, SIFT, and SIFT-mean shift, respectively, take 0.5, 0.8, and 0.9 s (in average) on a PC with a 1.5 GHz Intel(R) Pentium(R) CPU and MatLab implementation to process one image frame. It is an issue that these three tracking algorithm have not been able to achieve real-time performance. This issue will be dealt with once the implementation can be optimized.

## 5. Conclusions and future work

A solution to enhance the performance of classical mean shift object tracking has been presented. This work integrated the outcomes of SIFT feature correspondence and mean shift tracking. An expectation–maximization algorithm was proposed to optimize the probability function for a better similarity search. Experiments verified that the proposed method could produce better solutions in object tracking of different scenarios.

In future work, the research attempt is to investigate the convergence property of the proposed framework. This investigation may help enhance the proposed algorithm for efficiency purposes. In addition, this proposed algorithm needs to be comprehensively evaluated in a wider database. Currently, this paper suggests that, although the tracking results are promising in certain situations, further development and more evaluation is anticipated in severe image clutters and occlusions.

## Acknowledgments

The authors thank the anonymous reviewers for constructive comments that help improve the quality of this manuscript. Part of this work was conducted when the first author worked in the University of Essex, United Kingdom.

## References

[1] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Trans. Syst. Man Cyber.-C 34 (3) (2004) 334–352.
[2] G. Welch, G. Bishop, Scaat: incremental tracking with incomplete information, in: SIGGRAPH'97: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1997, pp. 333–344.
[3] M. Isard, A. Blake, Condensation—conditional density propagation for visual tracking, Int. J. Comput. Vis. (IJCV) 29 (1) (1998) 5–28.
[4] K. Choo, D. Fleet, People tracking using hybrid monte carlo filtering, in: Proceedings of the International Conference on Computer Vision, 2001, pp. II: 321–328.
[5] D. Reid, An algorithm for tracking multiple targets, IEEE Trans. Auto. Control 24 (6) (1979) 843–854.
[6] I. Cox, S. Hingorani, An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 18 (2) (1996) 138–150.
[7] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 25 (5) (2003) 564–577.
[8] A. Jepson, D. Fleet, T. El Maraghi, Robust online appearance models for visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 25 (10) (2003) 1296–1311.
[9] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.
[10] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: CVPR, 2000, pp. 142–151.
[11] Y. Cai, N. de Freitas, J. Little, Robust visual tracking for multiple targets, in: 9th European Conference on Computer Vision, 2006, pp. 107–118.
[12] D. Ross, J. Lim, M. Yang, Adaptive probabilistic visual tracking with incremental subspace update, in: Proceeding of the Eighth European Conference on Computer Vision, 2004, pp. 215-227.
[13] D. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the International Conference on Computer Vision ICCV, Corfu, 1999, pp. 1150–1157.
[14] L. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[15] H. Barlow, Single units and sensation: a neuron doctrine for perceptual psychology, Perception 1 (1972) 371–394.
[16] S. Ullman, The Interpretation of Visual Motion, MIT Press, Cambridge, MA, 1979.
[17] J. MacCormick, M. Isard, Partitioned sampling, articulated objects, and interface-quality hand tracking, in: Proceedings of the European Conference on Computer Vision, 2000, pp. 390–395.
[18] J. Sullivan, A. Blake, M. Isard, J. MacCormick, Object localization by bayesian correlation, in: Proceedings of the Seventh International Conference on Computer Vision, 1999, pp. 1068–1075.
[19] C. Yang, R. Duraiswami, L. Davis, Efficient mean-shift tracking via a new similarity measure, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, 2005, pp. 176–183.
[20] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (8) (1995) 790–799.
[21] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.
[22] M. Fashing, C. Tomasi, Mean shift is a bound optimization, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 471–474.
[23] D. Lowe, Robust model-based motion tracking through the integration of search and estimation, Int. J. Comput. Vis 8 (1992) 113–122.
[24] P. Wunsch, G. Hirzinger, Real-time visual tracking of 3D objects with dynamic handling of occlusion, in: Proceedings of 97 International Conference on Robotics and Automation, 1997, pp. 2868–2879.
[25] C. Harris, Geometry from Visual Motion in Active Vision, MIT Press, Cambridge, MA, USA, 1993.
[26] D. Gavrila, L. Davis, 3D model-based tracking of humans in action: a multi-view approach, in: Proceedings of the Computer Vision and Pattern Recognition, 1996, pp. 73-80.
[27] B. Schunck, B. Horn, Determining optical flow, in: DARPA81, 1981, pp. 144-156.
[28] B. Schunck, The image flow constraint equation, Comput. Vis. Graph. Image Process. 35 (1) (1986) 20–46.
[29] J. Shi, C. Tomasi, Good features to track, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994, pp. 593–600.
[30] E. Parzen, On the estimation of a probability density function and mode, Ann. Math. Statist. 33 (1962) 1065–1076.
[31] A. Elgammal, R. Duraiswami, L. Davis, Probabilistic tracking in joint feature-spatial spaces, in: CVPR03, 2003, pp. I: 781–788.
[32] T. Lindeberg, Scale-space theory: a basic tool for analysing structures at different scales, J. Appl. Statist. 2 (2) (1994) 224–270.
[33] P. Rousseeuw, Least median of squares regression, J. Am. Statist. Assoc. 79 (388) (1984) 871–880.
[34] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Statist. Soc., B 39 (1) (1977) 1–38.
[35] M. Carreira-Perpinan, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 767–776.
[36] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), South Carolina, 2000, pp. 142–149.