Adaptive Fusion Of Particle Filtering And Spatio-Temporal Motion Energy For Human Tracking

Huiyu Zhou[a,*], Minrui Fei[b], Abdul Sadka[c], Yi Zhang[d], Xuelong Li[e]

[a] *School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, BT3 9DT, United Kingdom*
[b] *School of Mechatronics Engineering and Automation, Shanghai University, China. Email: mrfei@staff.shu.edu.cn.*
[c] *School of Engineering and Design, Brunel University, United Kingdom. Email: Abdul.Sadka@brunel.ac.uk.*
[d] *Institute of Automation, Chongqing University of Post and Telecommunication, China. Email: zhangyi@cqupt.edu.cn.*
[e] *Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China. E-mail: xuelong_li@opt.ac.cn.*

Abstract

Object tracking is an active research area nowadays due to its importance in human computer interface, teleconferencing and video surveillance. However, reliable tracking of objects in the presence of occlusions, pose and illumination changes is still a challenging topic. In this paper, we introduce a novel tracking approach that fuses two cues namely colour and spatio-temporal motion energy within a particle filter based framework. We conduct a measure of coherent motion over two image frames, which reveals the spatio-temporal dynamics of the target. At the same time, the importance of both colour and motion energy cues is determined in the stage of reliability evaluation. This determination helps maintain the performance of the tracking system against abrupt appearance changes. Experimental results demonstrate that the proposed method outperforms the other state of the art techniques in the used test datasets.

Keywords: Computer vision, object tracking, occlusion, colour, motion energy.

*Corresponding author. Tel: +44-28-90971753; fax: +44-28-90971702.
Email address: h.zhou@ecit.qub.ac.uk (Huiyu Zhou)

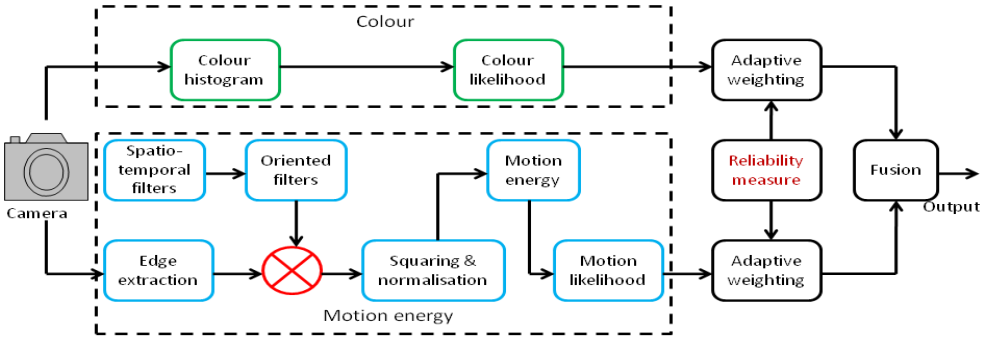


Figure 1: Flowchart of proposed spatio-temporal based particle filter tracking system.

1. Introduction

Object tracking can be defined as a process of establishing temporal coherent correlations between image features over consecutive frames according to their shape, appearance and distance information [1, 2, 3, 4, 5, 6]. Applications of object tracking have been commonly found in video surveillance [7], sports analysis [8], human motion analysis [9] and human-computer interface [10],[11],[12]. In particular, research studies on video surveillance applications mainly address challenging issues such as low quality images, illumination variations and cluttered backgrounds [13],[14],[15],[16]. Interesting scenarios reported in these studies consist of traffic monitoring, gate access, crowd control and transportation security [17, 18].

A working tracking system normally contains two key modules: prediction and correction. In the prediction module, a new state is obtained using the estimates made when people analyse the previous image frame, and a hypothesis is generated using a dynamic model. The correction module introduces an improved measurement, obtained at the current image frame, which refines the prediction. This new measurement is used in the next step as a new prediction. The system then assigns trust (or confidence) values to the prediction and the observation, respectively. The final estimation (or track) will be a compromise between the two locations which may or may not be close to the correct location. Nevertheless, it is hard to determine which resource is more trustable than the others in a challenging circumstance. This is mainly due to the appearance change of the subject to be tracked, or distraction because of the background clutters.

To deal with this problem, a tracking system needs reliable observations that are invariant to rotations, partial occlusions and pose changes.

For example, local features or salient points, e.g. colours, textures, shapes or gradients, are commonly used in classical tracking systems for reducing ambiguity. To handle complicated situations such as crowds, people have applied corner/edge detection and optical flow calculation onto the corresponding stage in order to improve the discriminative capability. Furthermore, multiple visual cues or features have been fused so as to maintain the consistency of the observations. For example, colour cues have been integrated with shape information for human tracking in real applications.

However, empirical evidence shows that this integration scheme can fail to work in the case where subjects get dresses in similar colours, and background clutters may cause excessive uncertainty and ambiguity in shape discrimination. To handle this problem, Adam et al. [19] proposed to use patches to describe an object, and then compared their histograms over the image frames. Babenko et al. [12] introduced an adaptive appearance model that embedded a discriminate classifier in an on-line manner to separate the object from the background. A low dimensional eigenbasis representation has been learnt, based on incremental principal component analysis algorithms [11]. In spite of promising outcomes in tracking accuracy, local feature based tracking systems have witnessed increasing challenges in practice (e.g. [20]). First of all, local feature based tracking systems require a selection of specific features extracted from the images. This subjective selection mechanism is operator dependent and may not effectively work in different environments. Secondly, it is a big challenge for one to estimate the centroid or velocity of a moving subject in the presence of occlusions.

In this paper, we introduce an adaptive feature fusion based tracking system to handle the problem of pose/illumination changes and occlusions. This scheme, despite being in its early stage of development, can be applied to event recognition and retrieval in the future incorporating the approaches such as [21]. Combining with the probabilistic data association algorithm e.g. [22], the proposed strategy can also be extended to the case of multiple person tracking. Fig. 1 shows the flowchart of the proposed system. We here propose to use an appropriate image representation, which refers to an effective fusion of local features (i.e. colour histograms) and the spatio-temporal motion energy [23]. The spatio-temporal representation can be used to capture the spatial appearance and dynamics of visual spacetime with certain immunity to appearance changes or clutters (this is different from the case reported in [24] that used colour and orientation cues). More importantly, when one of the fused features experiences fails, the other feature can be used to keep the systematic consistency and hence shows resilience to the complicated situations.

Compared to similar work such as [20],[25],[26], which take advantage of oriented energy features, multi-label and similar strategies, our major contributions are three-fold: (1) An adaptive fusion of the likelihoods associated with different features is presented in this paper. To be precise, motion energy cues are combined with colour histogram information within a particle filter based framework. It employs a reliability measure to determine the importance of colour and motion energy cues in the state estimation and updating, based on the historic colour or motion estimations. (2) The characteristics of the proposed scheme is investigated in terms of effectiveness and computational complexity. (3) A comprehensive evaluation of the proposed approach is performed against the other state of the art techniques.

This paper is organized as follows: Section 2 introduces the related work. Section 3 presents the proposed Spatio-tempOral Motion Energy Particle Filter (SomePF) tracking algorithm. This is followed by the description of the experimental work in Section 4. Finally, conclusions and future work are drawn in Section 5.

2. Related work

Prominent tracking systems developed to date include the Kalman filter [27] and the Condensation algorithm (also namely particle filter) [28]. Other existing techniques include mean shift [29], optical flow [30],[31], multiple hypothesis tracking [32],[33], Bayesian networks [34] and hidden Markov models [35].

Kalman filter has a long history to be intensively used in the community [27]. It is a state estimate method, based on linear dynamical systems that are perturbed by Gaussian noise. In most cases, a uni-model probability distribution is used in the state estimate. To release the linear assumption, a non-linear version, called extended Kalman filter (EKF), has been commonly used in object tracking [36]. However, the EKF also experiences some major issues in practice, one of which is that without the assumption of "static noise", the estimated covariance matrix becomes unstable and hence causes the estimations drift away. Similar work like the iterated extended KF and unscented KF has also been reported in the community.

Particle filter is a conditional density propagation method that can be used to deal with non-Gaussian distributions and multi-modality cases [28]. This method allows a posterior distribution, estimated in the previous image frame, to be sampled with a set of particles. These particles are then propagated iteratively to successive frames using continuously updated observations and a prediction model. Particle filters have been popularly used

to handle various tracking problems, e.g. [37],[38],[39]. However, the performance of particle filters can degrade as the dimensionality of the state space increases, and the support of the likelihood decreases (e.g. less sample numbers). One of the effective solutions is to combine a variational approximation with efficient importance sampling to achieve tracking recursion [39].

Mean shift is a technique of locating the maxima of a density function using the samples from that function [40]. This involves an iterative procedure, which starts with an initial guess and stops when the difference between two consecutive estimates is smaller than a pre-defined threshold. A kernel function is used to determine the weight of neighbouring points for re-estimation of the mean value. Mean shift has been used with colour histograms to find the peak of a confidence map for locating an object's position [29]. Yilmaz [41] introduced an asymmetric kernel mean shift, in which the scale and orientation of the kernel adaptively change depending on the observations at each iteration. Freedman and Kisilev [42] presented a novel fast mean shift procedure, based on random sampling of the Kernel density estimates. Yeh and Hsu [43] reported a feature selection algorithm used within a mean shift scheme, which adopted the AdaBoost algorithm to select features that best compensate each other and determine their weights using likelihood estimations. A comprehensive evaluation has been conducted in [44], which shows that mean shift based trackers have better performance than the variant CAMShift tracker.

Optical flow based tracking systems assume brightness constancy over a short period. For example, the well-known Lucas-Kanade method considers that the displacement of the image contents between two consecutive frames is small and approximately constant within a neighborhood of an image point under consideration [30]. Shi and Tomasi [31] investigated two spatial gradient matrices that can be used to determine the quality of each corner feature. After good features have been determined, the Lucas-Kanade optical flow algorithm is then applied.

Stochastic tracking systems, to some degree, were overwhelmingly studied in the last decade. These systems usually model the similarity between two observations as random variables and then learn their probability distributions in a stochastically optimal manner. For example, Kwon and Lee [45] proposed a tracking framework that tracked a target by searching for the appropriate trackers in each frame. The system collected several samples of both the states of the target and the trackers during the sampling process. On the other hand, the Wang-Landau sampling methods was integrated with a Markov Chain Monte Carlo (MCMC) based tracking framework. This was achieved by applying the density-of-states term estimated by the

Wang-Landau sampling method to the acceptance ratio of MCMC to alleviate the motion smoothness constraint [46]. Schaap et al. [47] introduced a Bayesian tube tracking algorithm that incorporated a priori knowledge to enhance the tracking performance of tube structures. Cui et al. [48] reported a Monte Carlo tracker for tracking s single rolling leukocyte in vivo. A specilised sample-weighting criterion was applied to the tracker, cased on the leukocyte movement and intensity features. Cheng et al. [49] presented a vessel-tracking scheme that generated plenty of sampling paths to describe the complicated topology of the vascular structures with calcium depositions. A compiling and linking process was utilised to organise the sampling paths together to form the vessel segments that may belong to the same vessel tract.

In recent years, spatio-temporal representations gain increasing attention in object detection/tracking applications. Particularly, the use of orientation selective filters helps enhance visual tracking performance because they can capture the dynamic characteristics of motion patterns. Since these orientation selective filters report the correct direction of the object' motion, the tracker finds that it is straightforward to establish correspondences across frames. This feature allows the motion energy based trackers to be distinguished from the other state of the arts especially in the circumstance of rotations. For example, [50] shows that motion energy was used to consistently track objects in the crowd areas, whilst the feature based tracking systems strived to deal with the standard correspondence problem.

Lam and Shi [51] successfully applied motion energy neurons for active visual tracking of heading directions. In the meantime, it was reported that motion energy can be used as an indicator of motion direction [52]. Chen et al. [25] used spatio-temporal oriented energy for moving crowd detection. Cannons and Wildes [53],[20] proposed oriented energy features with a robust regression measure scheme. The oriented energy used in the approach encodes the local spatio-temporal structure at a specific orientation/scale level. However, evidence also shows that every MS based tracker cannot guarantee global optimality, and thereby falls into local maxima [54].

Xiang and Gong [55] presented a Dynamic Probabilistic Networks (DPNs) based approach for modelling the temporal and causal correlations among discrete events in order to achieve robust and holistic scene-level behaviour interpretation. Cui et al. [56] reported a new strategy to perform hierarchical event recognition using a sequential Monte Carlo method. In [57], a novel representation and recognition technique was presented for identifying movements, based on temporal templates and their dynamic matching in time.

3. Proposed spatio-temporal tracking algorithm

3.1. Colour based particle filtering

As an attempt to handle the complexity of object tracking, our proposed framework is intended to search for a global maxima that leads us to finding the best matches across two image frames through spatio-temporal analysis.

Our proposed tracking system is based on a particle filter paradigm. Particle filtering was originally used to address the tracking problem in clutters [28]. It is a technique for implementing a recursive Bayesian filter by Monte Carlo simulations. The goal of particle filtering is to compute the posterior density $p(\mathbf{X}_t|\mathbf{Z}_t)$, where the vector $\mathbf{X}_t$ denotes the state of the tracked object, $\mathbf{Z}_t$ is the observations $(\mathbf{z}_1, ..., \mathbf{z}_t)$ up to time t. In this paper, both the posterior density $p(\mathbf{X}_t|\mathbf{Z}_t)$ and the observation density $p(\mathbf{z}_t|\mathbf{X}_t)$ are presumably non-Gaussian although this is not necessary.

In the particle filter domain, the posterior density is approximated by a weighted particle set $\{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$ at each time t. Each particle $s_t^{(n)}$ is one hypothetical state of the object, weighted by a discrete sampling probability $\pi_t^{(n)} = p(\mathbf{z}_t|\mathbf{X}_t = s_t^{(n)})$, which is the probability of current observations being generated by the hypothetical state. In the meantime, $\sum_{n=1}^N \pi_t^{(n)} = 1$. The particle filtering scheme accommodates multiple state hypotheses simultaneously, and is capable of coping with short term occlusions, when object states gradually change with certain instability.

As demonstrated in the literature, colour distributions can be used as a target model, which preserves a partial identity in different environments [58],[24]. In this paper, we follow the convention used in [58] to present the colour based particle filter. Let the colour distributions be discretised into m-bins. The histograms are represented as $h(\mathbf{x}_i)$, where $\mathbf{x}_i$ indicates the corresponding bin of a location. Here, the histograms are computed in the RGB space with 8×8×8 bins. HSV can also be used (not here though), which is not sensitive to V components in the case of lighting changes. Let an elliptic area be a track, which has half major/minor axes $\mathbf{H}_x$ and $\mathbf{H}_y$.

To determine whether or not a pixel belongs to the background or get occluded by something else, a weighting function $k(r)$ is used to define the ownership of the pixel [58]: when the distance r (a normalised value) from the pixel to the region centre is larger than a pre-defined threshold, then the weight is zero. Otherwise, $k(r) = 1 - r^2$. To calculate the colour distribution $p_M = \{p_M^{(\alpha)}\}_{\alpha=1,...,m}$ at location $\mathbf{M}$, we have the probability of the colour η falling in a candidate region: $p_M^{(\alpha)} = f_0 \sum_{i=1}^m k\left(\frac{\|\mathbf{M}-\mathbf{x}_i\|}{h}\right) \delta[h(\mathbf{x}_i) - \eta]$,

where m is the number of pixels in the outlined region, δ is the Kronecker delta function, $h = \sqrt{\mathbf{H}_x^2 + \mathbf{H}_y^2}$ and the normalisation factor: $f_0 = 1/\sum_{i=1}^{m} k\left(\frac{\|\mathbf{M}-\mathbf{x}_i\|}{h}\right)$, which satisfies $\sum_{\alpha=1}^{m} p_M^{(\alpha)} = 1$.

In a particle filter, new observations are continuously acquired and the state vector is dynamically updated. To evaluate the changes in the observations, we apply a similarity measure for two distributions $p(\alpha)$ and $q(\alpha)$ (e.g. colour histograms) using the Bhattacharyya coefficient [59] that is an absolute similarity measure and needs no bias correction: $\rho[p,q] = \int \sqrt{p(\alpha)q(\alpha)}d\alpha = \sum_{\alpha=1}^{m} \sqrt{p^{(\alpha)}q^{(\alpha)}}$. Therefore, the Bhattacharyya distance between the two distributions can be defined as

$$d = \sqrt{1 - \rho[p,q]}, \tag{1}$$

which suggests that as ρ becomes larger, the two distributions are more similar.

When generating random samples from the prior distribution, we represent a sample as an ellipse with the following vectorised form: $\mathbf{s} = \{x, y, \dot{x}, \dot{y}, \mathbf{H}_x, \mathbf{H}_y, \dot{h}\}$, where ($x$,$y$) are the coordinates of the location of the ellipse, ($\dot{x}$,$\dot{y}$) the velocity and $\dot{h}$ the scale change. To predict the location of a sample in successive image frames, a first order dynamic model is used to describe the motion (of constant velocity and scale change): $\mathbf{s}_t = \mathbf{A}\mathbf{s}_{t-1} + \mathbf{w}_{t-1}$, where $\mathbf{A}$ is the deterministic vector and $\mathbf{w}_{t-1}$ is a Gaussian random variable.

Colour distributions play a key role in the sampling stage for similarity check. Normally, large weights are given to the region, which has a similar colour distribution to that of the target model. Here, the target model refer to the colour distribution. Hence,

$$\pi^{(n)} == \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1 - \rho[p_{\mathbf{s}^{(n)}}, q]}{2\sigma^2}\right), \tag{2}$$

where σ is a Gaussian variance and $\mathbf{s}^{(n)}$ is a dynamic model at time n. This ensures that the samples of higher weights be selected more often than those of lower weights. Finally, the best state vector is determined using a Monte Carlo approximation of the expectation $\mathbb{E}(\mathbf{X}_t|\mathbf{Z}_t)$ over all the particles.

3.2. Spatio-temporal motion energy

Due to the smoothness constraint, object tracking can be made relatively easy if the motion of the object to be tracked in successive frames is well predicted and its dynamics can be properly modeled. The observation likelihood of the object can be formed using size, motion and appearance cues.

Size and appearance descriptors can be dramatically affected in the presence of view line and illumination changes [11]. As a result, corner/edge/blob and shape features become unreliable during the tracking period.

Motion estimation has been studied as a fundamental problem for many years and significant progress has been achieved [23],[51],[20]. An effective framework for representing the spatial appearance and motion characteristics is the use of motion energy (also namely oriented energy in [53],[20]). Motion energy is indeed a 3-D component (x,y,t), in which x and y are two spatial dimensions and t is the temporal dimension. The energy value is produced as the filter response of orientation selective bandpass filters, which can be applied to each single frame in a video sequence. It is worthy to point out that an edge map is here used as a stimulus in order to generate the filter response.

Motion energy has been demonstrated as a well-formed feature for visual tracking applications due to its merits as follows [20]: (1) It is a comprehensive descriptor that encompasses appearance and dynamics. A tracker based on this descriptor has a potential to effectively work against the impact of background clutters. (2) This feature appears to be robust to illumination changes. (3) Motion energy is formed in the context of multi-scaling, resulting in a well defined structure with rich textures. (4) The representation can be implemented using linear and pixel-wise non-linear operations and the tracker can run in real-time.

Motion energy can be derived using properly tuned three-dimensional ((x,y,t)) Gaussian second derivative filters, $G_2(\theta, \gamma)$, and the corresponding Hilbert transforms, $F_2(\theta, \gamma)$, where θ refers to the three-dimensional at which filtering is being performed, and γ is the scale of a Gaussian pyramid [23]. The second derivative filters are chosen because of its easy implementation so that the motion can be modeled as a local nonlinear process such as flicker detection or extraction of unsigned contrasts. This is followed by a spatio-temporal correlation of the signals [60]. Higher order derivative filters are possibly used, and evidence suggests that the higher order derivative filters can account for motion characteristics in specific cases, similar to the lower order cases [61].

Motion energy is the filter response in the format of pixel-wise rectification and summation as follows:

$$E(\mathbf{u}; \theta, \gamma) = [G_2(\theta, \gamma) * I(\mathbf{u})]^2 + [F_2(\theta, \gamma) * I(\mathbf{u})]^2, \tag{3}$$

where $\mathbf{u} = (x, y, t)$ are the spatio-temporal image coordinates over time, I is an image, and $*$ indicates the convolution operator. Eq. (3) uses a

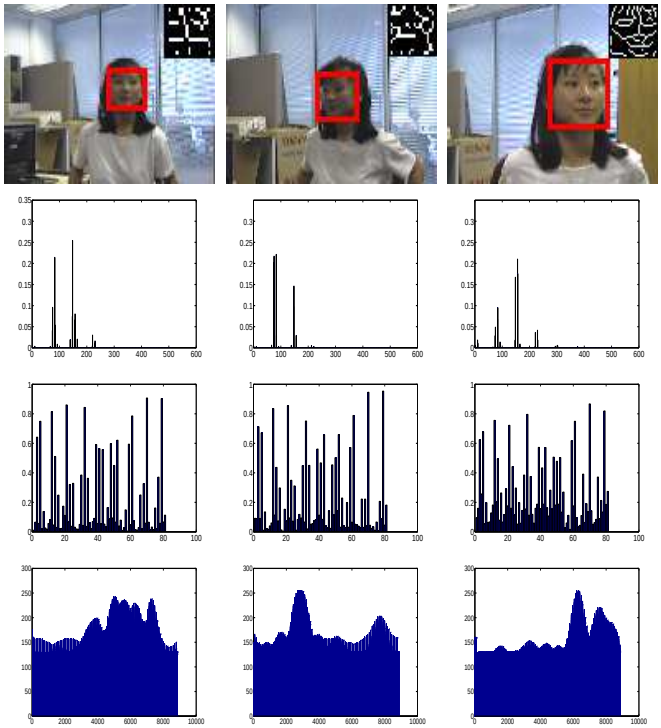


Figure 2: Illustrations of different features for out-of-plane rotation (col. 1: facing forward, col. 2: left sideways, col. 3: right sideways). Row 1 - original images with tracking results and the Canny edges (right-up corner); row 2 - $8 \times 8 \times 8$ RGB histograms; row 3 - HOG histograms; row 4 - bar chart plots of motion energy features corresponding to the tracks (standard deviation = 0.2 deg, frequency of carrier = 1.1 cfd), where x-axis indicates feature index and y-axis refers to feature values.

modulation form so that energy can be extracted within a spatio-temporal frequency band. The filtering response is sensitive to the direction of motion but insensitive to the sign of the image contrast. This characteristic provides phase invariance, with sign of contrast as an extreme example.

Fig. 2 illustrates the features extracted from the images of different motion patterns. These features consist of motion energy and two popularly used features in the literature: color histograms and histograms of orientation gradients (HOG). The out-of-plane rotation produces rapid changes to the image representations. As a result, we observe that (1) the motion energy scheme leads to much denser distributions and richer details than the color and edge based trackers, and (2) the motion energy scheme provides the estimation of facial direction but the color and edge based trackers cannot do so. In this example, the highest peaks on the 4th row of Fig. 2 correctly indicate the facial directions of the corresponding images on the 1st row (assuming that the middle point of the range along the x-axis refers to the direction of facing forward). The capability of correctly reporting directional information enables the motion energy based system to have better tracking performance than the other state of the art techniques (also see the experimental work).

To allow different elements to have equal contributions, whilst reducing estimation errors, we perform a pixel-wise normalisation as a further processing step:

$$\hat{E}(\mathbf{u};\theta,\gamma) = \frac{E(\mathbf{u};\theta,\gamma)}{\sum_{\tilde{\gamma}}\sum_{\tilde{\theta}} E(\mathbf{u};\tilde{\theta},\tilde{\gamma}) + \nu\bar{\sigma}_n}, \tag{4}$$

where ν is a constant (empirical values fall in the range of (1,10)) that can be used to modulate the contribution of $\bar{\sigma}_n$, the average variance of the entire image. To obtain a reasonable $\bar{\sigma}_n$, we use such a technique as: the overall pixels of the input image are grouped using a K-means clustering method [62], according to intensity levels. Each group is represented as a linear slope and spatial extent. The gray levels within each group are of a variance $\hat{\sigma}_i$. $\bar{\sigma}_n$ is the mean of $\hat{\sigma}_i$.

Our approach, shown as Eq. (4), can be used to handle different image noise (i.e. $\bar{\sigma}_n$). This is a better approach than the one introduced in [20], which simply applied a constant τ to describe the influence of non-stationary image noise. The tilde symbol in Eq. (4) is used here to denote different scales and orientations, used in the filtering stage. Such a normalisation helps regulating the changes in motion speeds and image contrasts, where abrupt changes in one or both of the components will be constrained as a result.

3.3. Integrating spatio-temporal motion energy with colour cues

We now investigate the fusion of colour distributions and spatio-temporal motion energy. To combine these two different items into a single form, one feasible way is to facilitate an appropriate combination of two distributions (or likelihoods). For this purpose, we first calculate the likelihood of each cue using the distance from the model histograms defined in Eq. (1) as follows:

$$p_j(\mathbf{z}_t|\mathbf{X}_t) \propto \exp\left(-\frac{d(f_j(\mathbf{X}_t), q_{j,t})}{\sigma_0}\right), \tag{5}$$

where $q_{j,t}$ is the distribution of each cue used as a template for correspondence. The histogram $f_j(\mathbf{X}_t)$ (item index j = 1,2 indicating the contributions from colour and motion energy respectively) against the state $\mathbf{X}_t$ is computed over the overall pixels of the observation vector $\mathbf{z}_t$. The intensity histogram $f_{\hat{E}}(\mathbf{X}_t)$ (or named $f_2(\mathbf{X}_t)$) of motion energy features $\hat{E}(\mathbf{u};\theta,\gamma)$ (outcome of Eq. (4) that is applied to the candidate area) is experimentally formed with 216 bins, followed by a standard normalisation step. Note that these bins are formed using 80 spatial samples and 100 temporal samples in the filter that has a half-width of 2 degrees. The fast and slow temporal filters have the widths of 6 and 9, respectively. Furthermore, the exponential form as shown in Eq. (5) is adopted so that a smooth likelihood can be formulated for the state estimation. The variance σ_0 is determined experimentally, and it can be influenced by either motion cues or colours.

Once the individual likelihoods have been obtained, the conditional probability is then produced using linear combinations of all the cues [63]:

$$p(\mathbf{z}_t|\mathbf{X}_t) = \sum_{j=1}^{2} \beta_{j,t} p_j(\mathbf{z}_t|\mathbf{X}_t), \tag{6}$$

where $\beta_{j,t}$ is a mixture coefficient that is constrained by

$$\begin{cases} \sum_{j=1}^{2} \beta_{j,t} = 1, \\ \beta_{j,t} \propto \beta_{j,t-1} p(\mathbf{X}_{t-1}|\mathbf{z}_{t-1}). \end{cases} \tag{7}$$

We also have a prior as $p(\mathbf{X}_0) \equiv p(\mathbf{X}_0|\mathbf{z}_0)$. The summation used in Eq. (6) makes the probability estimation less sensitive to inconsistent measurements acquired in specific circumstances, e.g. one of the cues is suddenly unavailable [22]. This strategy is motivated by the fact that humans are capable of perceiving scenes through additive collections of visual features weighted by reliability measures [64].

Using Bayes' rule, we can decompose the posterior $p(\mathbf{X}_{t-1}|\mathbf{z}_{t-1})$ to the following form:

$$p(\mathbf{X}_{t-1}|\mathbf{z}_{t-1}) = \frac{p(\mathbf{z}_{t-1}|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{z}_{t-2})}{p(\mathbf{z}_{t-1}|\mathbf{z}_{t-2})}, \tag{8}$$

where the likelihood $p(\mathbf{z}_{t-1}|\mathbf{X}_{t-1})$ can be estimated using the measurement model described above. Meanwhile, the evidence is represented by

$$p(\mathbf{z}_{t-1}|\mathbf{z}_{t-2}) = \int p(\mathbf{z}_{t-1}|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{z}_{t-2})d\mathbf{X}_{t-1}. \tag{9}$$

This evidence defines the "prior predictive distribution" of $\mathbf{z}_{t-2}$, given the systematic model. It can be set to a constant if the prior statistics have been established, e.g. using the mean of the measurements within the investigated region. It is worth pointing out that, if $\mathbf{z}_{t-1}$ and $\mathbf{z}_{t-2}$ are significantly different, then the resulting posterior may lead the tracker to an incorrect match. This possibility can be effectively reduced using the combination of multiple cues, which can reach a compromise between two estimations and therefore minimise "drifts". This will be further justified in our experiments. The prior is

$$p(\mathbf{X}_{t-1}|\mathbf{z}_{t-2}) = \int p(\mathbf{X}_{t-1}|\mathbf{X}_{t-2})p(\mathbf{X}_{t-2}|\mathbf{z}_{t-2})d\mathbf{X}_{t-2}. \tag{10}$$

To simplify the computation, the prior can be approximated as

$$p(\mathbf{X}_{t-1}|\mathbf{z}_{t-2}) \approx \frac{1}{(2\pi\sigma_z^2)^{N_z/2}} \exp\left(-\frac{1}{2\sigma_z^2}\int (\mathbf{x}_z - \mu_z)^2 d\mathbf{x}_z\right). \tag{11}$$

where $\mathbf{x}_z$ is the target location at $(t-1)$, μ_z and σ_z are the mean and variance of the distance between each sample and the weight centre of N_z samples, respectively.

To discard the samples with low weights that cause numerical instability, whilst avoiding degeneracy, we apply a resampling scheme (bootstrapping, or Monte Carlo simulations). In the resampling stage, we use a resampling function defined as follows:

$$\phi_t^i = \sum_{j=1}^{2} \lambda_{j,t} p_j(\mathbf{z}_t|\mathbf{X}_t^i), \tag{12}$$

where random samples $i = 1,...,N_s$, $\mathbf{X}_t^i$ indicate the state vectors due to i samples, and the scaling factor

$$\lambda_{j,t} = \begin{cases} \varphi_{j,t}, & if \ \ \varphi_{j,t} > T_0, \\ T_0, & otherwise. \end{cases} \tag{13}$$

In [24], $\varphi_{j,t}$ refer to thresholds and T_0 is the minimum number of the particles resampled from each cue. In our approach, however, T_0 is dynamically determined according to the following function:

$$T_0 = \exp\left(-\frac{\varsigma^2}{2\sigma_T^2}\right), \tag{14}$$

where $\varsigma = \mu_c(1-\Phi_c) + (1-\mu_c)(1-\Phi_m)$ ($\mu_c$is determined in a training stage), and the variance σ_T is experimentally defined. Functions Φ_c and Φ_m, accounting for similarity measures from color and motion energy features respectively, are individually calculated using the Mahalanobis distance between each current sample area and the previous detection. Note that Φ_c and Φ_m must be normalised so that ς is a positive real number. In terms of colour features, a cumulative histogram is generated with $8 \times 8 \times 8$ bins.

On the other hand, the weight $\{\lambda_{j,t}\}_{j=1}^{2}$ is also normalised so that ϕ_t^i fall in the range of [0,1]. This normalisation process can help balance the contributions of different cues in a circumstance such as one of them is missing. If the number of the samples is large enough (> 200), the posterior $p(\mathbf{X}_t|\mathbf{z}_t) \approx \frac{1}{N_s}\sum_{j=1}^{2}\delta(\mathbf{X})^j$.

Algorithm 1 shows the resampling scheme in the proposed tracking algorithm.

Algorithm 1 Resampling process used in the tracking system.

1: Initialisation
2: Random draw $R \sim \mathcal{U}[0,1)$
3: **for** i = 1,...,N_s **do**
4: Calculate $\lambda_{j,t}$ using Eqs. (13)-(14)
5: Apply Eq. (12)
6: $\mathcal{O}_i \leftarrow [\phi_t^i N_s/\phi_t^{N_s} - R] + 1$
7: **end for**

3.4. Adaptive weighting of multiple cues

In Eq. (6), the weight $\beta_{j,t}$ is determined using the reliability measurement of each cue (i.e. colour or motion energy distribution). Fig. 3 illustrates the adapted contributions of colour and motion energy histograms in a face tracking scenario. Maggio et al. [24] proposed to define the impact of each feature based on individual spatial uncertainty, stemming from the study reported in [64]. The spatial uncertainty analysis is performed using

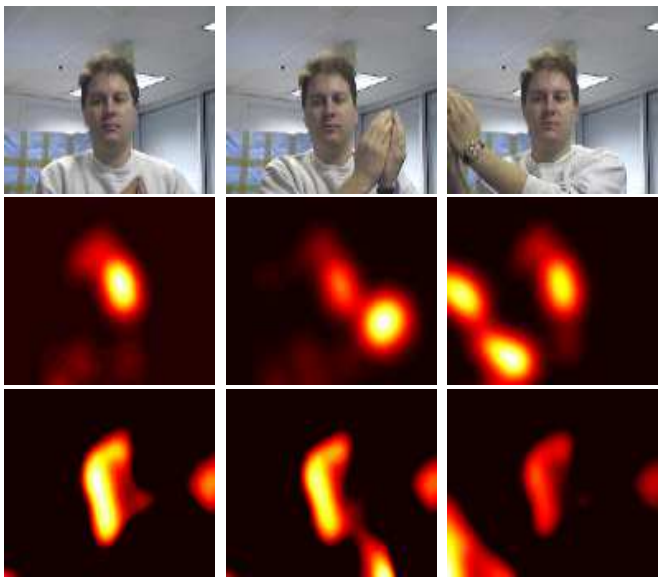


Figure 3: Comparison between the candidate likelihood of colour and motion energy histograms in a face tracking scenario. Row 1: original images. Row 2: Spatial distribution of the colour likelihood. Row 3: Spatial distribution of the calculated motion energy. In row 3, a reliability score is applied to adaptively assign a larger weight to the motion energy than the colour because the former provides consistent measurements of the face (standard deviation = 0.2 deg, frequency of carrier = 1.1 cfd).

the eigenvalues of the covariance matrix of the particles $C_{j,t}$ (we follow the symbol convention of [24]) weighted by the likelihood that is computed for each cue/feature j at time t. Given $\mathbf{X} \rightarrow (u, v)$, the covariance matrix is defined as:

$$\tilde{C}_{j,t} = \begin{bmatrix} \frac{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)(u^i-\hat{u})^2}{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)} & \frac{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)(u^i-\hat{u})(v^i-\hat{v})}{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)} \\ \frac{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)(u^i-\hat{u})(v^i-\hat{v})}{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)} & \frac{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)(v^i-\hat{v})^2}{\sum_{i=1}^{N_s} \mathcal{N}(s^i,v^i)} \end{bmatrix}, \quad (15)$$

where $\mathcal{N}(\cdot, \cdot)$ is a normal distribution of the samples, s and v are the mean and standard deviation respectively.

The uncertainty is determined as: $U_{j,t} = \sqrt[D]{det(\tilde{C}_{j,t})}$, where D is the dimensionality of the state space (5 is used here). The determinant $det(\cdot)$ is used instead of the summation of the eigenvalues for the stability purpose. Finally, the reliability score is defined as: $\psi_{j,t} = \frac{1}{U_{j,t}}$. Each feature (or cue) has the importance proportional to its uncertainty measure.

In spite of its success in some tracking applications, this reliability analysis has to handle difficult problems such as consistency and computational complexity. In the presence of clutters, for example, the dominant eigenvalues of the clutters are very close to those of the targets [65]. Consequently, there is a need to distinguish between the eigenvalues of the clutters and those of the correct target using effective post-processing techniques such as sub-windowing in which thresholding can be applied in individual neighboring areas.

We now re-visit Eq. (15). It is quite often to observe that the covariance matrix has relatively large values sitting on the diagonal elements, and the off-diagonal elements have small values. In this case, the mean and deviation differences are modulated by the "weights" $\mathcal{N}(\cdot, \cdot)$. A successful balance between colour and motion energy cues is fully dependent on individual errors (or residuals). Therefore, our approach uses a democratic integration method which is significantly different from [24], which mainly looked at the coherence between the colour or orientation cue at two neighbouring instances. We actually present a re-weighting scheme, where each feature/cue is associated with a score based on the residual between the individual cues and the average saliency value. We first compute the score at time t for each cue, using the residual between the predicted location of each cue and the final settlement of the fused cues. This residual is measured as an Euclidean distance between the two centroids of the locations.

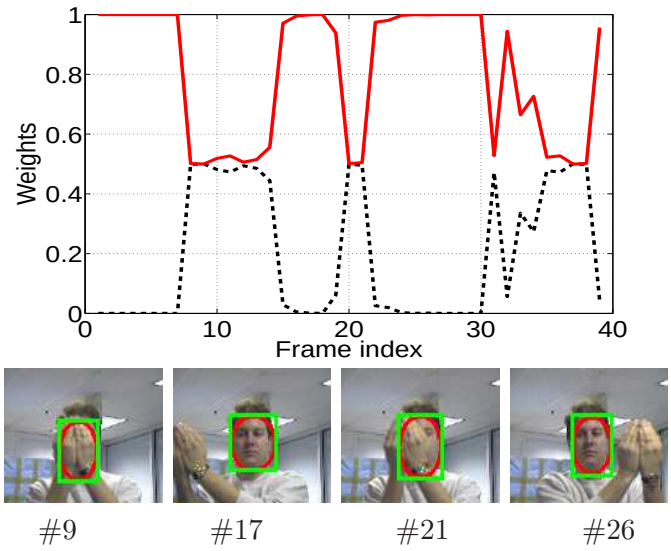


Figure 4: The evolution of the cue weights for the proposed tracking algorithm. Upper row: Red solid line - motion energy cue, and black dash line - colour cue. At frames 9 and 21, the motion energy cue share similar weights with the colour cue. In these frames, the hand is actually overlaid over the face and hence the two cues equally contribute to the estimation. At frames 17 and 26, the motion energy cue works well whilst the colour cue becomes unreliable due to the distraction of the hand nearby. Lower row: Ellipse - tracked targets, and rectangle - the regions used for computing motion energy.

Table 1: Characteristics of the testing databases, and H - head sequences and P - pedestrian sequences.

Sequences	Resolution	Frame rates (fps)	Descriptions
H1-H4	128×96	25	Clutter, occlusion, scale change
H5	320×240	25	Fast moving, scale change
P1-P3	360×288	25	Clutter, occlusion, scale change

If the evaluation period is too short, there may be insufficient historic estimates that can be used to generate a good prediction. To cope with this problem, we take into account a mean estimate of the residual measurements $\epsilon_{j,(t-l-1):(t-1)}$ over the previous l image frames (in our experiments, $l = 10$). Our approach does not apply extra computational efforts to the system. In fact, it chooses optimal weights according to the historic estimates. The score is then determined using a hyperbolic tangent function that is positive, bounded and symmetric:

$$\mu_{j,t} = \frac{\tanh(-c_1 \sum \epsilon_{j,(t-l-1):(t-1)}/l + c_2) + 1}{2}, \tag{16}$$

where the slope c_1 and the cut-off c_2 are experimentally defined ($c_1 = 20\ pixel^{-1}$ and $c_2 = 3$). The weights used in Eq. (13) are dynamically adapted through a leaky integrator:

$$\varphi_{j,t} = c_3 \mu_{j,t} + (1 - c_3) \sum \varphi_{j,(t-l-1):(t-1)}/l, \tag{17}$$

where c_3 is a scalar, which tunes the changing rate of φ according to individual experimental environments [66] and can be statistically determined in a training stage. We use the leaky integrator because (1) this integrator is easy to be implemented, (2) it is based on temporal correlation and (3) the integrator applies empirical and on-line estimator and therefore has certain capability to cope with noisy estimates. Without this leaky integrator, the weights would not be properly updated. To constrain the distribution's defusion, we re-normalise $\varphi_{j,t}$ and $\mu_{j,t}$ during the updating process.

In summary, to adaptively combine the measurements of colour and motion energy, we have the following process: (1) Eq. (16) is used to calculate a reliability score $\mu_{j,t}$, and the weight $\beta_{j,t} \propto c_w \mu_{j,t}$ (subject to a scalar c_w

Algorithm 2 The proposed tracking algorithm.

Initialising the Particle Filter: $\mathbf{X_t} = \bar{\mathbf{X}}_t = 0$
for m = 1,...,M **do**
  Sample state $\mathbf{x}_t^m \sim p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1}^m)$ using Eq. (11)
  Weight $W_t^m = p_j(\mathbf{z}_t|\mathbf{x}_t^m)$ by Eq. (5)
  $\bar{\mathbf{X}}_t = \bar{\mathbf{X}}_t + < \mathbf{x}_t^m, \mathbf{W}_t^m >$
end for
for m = 1,...,M **do**
  Draw i with probability $\mathcal{O}_i$ by Algorithm 1
  Add $\mathbf{x}_t^i$ to $\mathbf{X}_t$
end for
Return $\mathbf{X}_t$

experimentally determined). (2) Eq. (17) is used to calculate the scaling factor $\lambda_{j,t}$, shown in Eq. (13). Algorithm 2 shows the proposed tracking algorithm. Fig. 4 reveals the evolution of the color and motion energy cues. It is noticed that in this example, despite severe occlusions, the proposed integration scheme successfully assigns low weights to the cues which are unreliable for tracking. For example, at frames 17 and 26, the colour of the hands is quite similar to that of the face. In this circumstance, the weight of the colour cue is zero and the weight of the motion energy cue is one, which are computed using Eq. (16). At frame 20, where the face is occluded by the hands, both the weights of the two cues equal to 0.5. This example clearly shows that our system can handle the distraction problem in the presence of occlusions.

3.5. Complexity analysis

In this subsection, the computational complexity of the proposed "SomePF" tracker is discussed from a theoretical point of view. In a particle filter, there are three major operations: particle generation, weight calculation and resampling. Among these operations, resampling is more important than the remainder as this is used to entail improved estimations by enhancing the areas of higher posterior probability. The convergence property of resampling dominates the computational complexity of the particle filter.

Let the algorithm for random resampling be systematic resampling (SR). Given the number of the particles used in the particle filter, the complexity of the SR algorithm is $O(\max(N_{pf},M_{pf}))$, where N_{pf} and M_{pf} are the input and output numbers of the particles used in the resampling. On the other hand, the net complexity of spatio-temporal motion energy computation is

$O(w_{me}h_{me})$, where w_{me} and h_{me} are the dimensions of the image frame. As a result, the computational complexity of the proposed "SomePF" tracking system is approximately of $O(\max(N_{pf}, M_{pf}) + w_{me}h_{me})$. In our experiments, the computation of motion energy takes 5-10% of the time cost of the proposed "SomePF" tracker. For example, given 100 particles for the particle filtering scheme and a computer of Intel(R) Core(TM) 2.5GHz CPU, our system can process 3 images per second in the Matlab environment, which is similar to that of a classical particle filter based tracker.

4. Experimental work

4.1. Experimental set-up

The proposed tracking algorithm is evaluated using a database consisting of eight videos, each of which has one or more subjects. Four videos with human faces come from a publicly accessible dataset (`http://www.ces.clemson.edu/~stb/research/headtracker/`), and the remainder is a part of an in-house dataset [24]. In all the test sequences, the illumination conditions continuously change. For convenience, we tabulate the characteristics of the testing dataset in Table 1, and exemplar images of each video are illustrated in Fig. 5.

Our fusion based tracking algorithm ("SomePF") is compared against (1) colour based particle filtering ("CPF"), (2) fusion of colour and shape based tracker ("CEPF"), where a shape is represented by an edge histogram and the fusion outcome O_f is subject to a fixed weight ω: $O_f = \omega h_{colour} + (1-\omega)h_{shape}$ (in the experiments ω is set to be 0.5), (3) spatial-temporal motion energy based particle filtering ("SPF") without colour elements, (4) incremental visual tracking ("IVT") [11], and (5) on-line multiple instance learning based tracking ("MilTrack") [67].

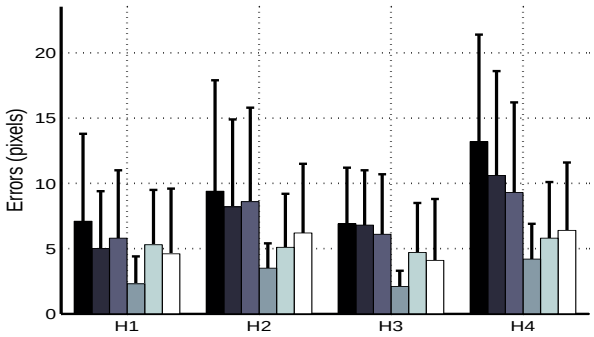
The entire evaluation consists of performance comparisons in the cases of distractions, occlusions, illumination changes and scaling. To ensure the experimental consistency, we run each experiment for 20 times, and the final statistics are the average values obtained over these 20 trials. To analyse the results, we use a score ϵ that describes the 2-D Euclidean distance between the detected position and the ground truth. The less the distance is, the better performance the system achieves. Since we focus on the case of single objects, there will not be any opportunity for identity switches. The score ϵ contains means and deviations.

Figure 5: Image samples of the eight test videos. Row 1: four sequences with faces (H1-H4); row 2: 2nd one is a sequence with faces (H5) and the remainder are pedestrian sequences (P1-P3).

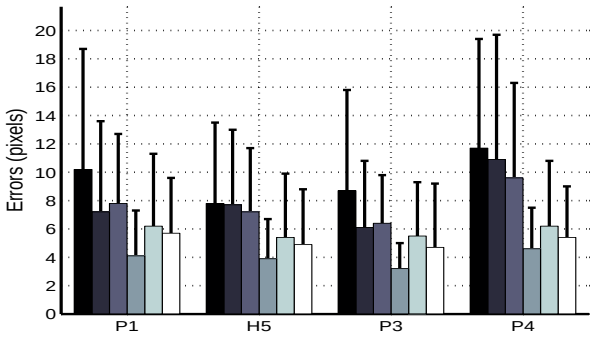
4.2. Performance comparison

First of all, we summarise the general performance of all the tracking approaches involved in the comparison and illustrate the statistical results in Fig. 6. In spite of occlusions and scale/illumination changes, the proposed "SomePF" algorithm has less errors than the other three algorithms. The proposed fusion based algorithm has sufficient capabilities to handle the case of occlusions and scale changes. This is because the "SomePF" approach takes the advantage of combining colour and motion energy cues: when one of them becomes unstable, the other one is used to provide reliable observations. We also observe that the outcomes of "CEPF" has better accuracy than "CPF". This attributes to the fact that the shape information used in "CEPF" helps generating discriminative distributions, which allow us to separate the target from its background. Both "IVT" and "MilTrack" better perform than the other methods except "SomePF". SPF has similar performance to CEPF in most of the video sequences.

We now take a closer look at the tracking results of different approaches. In the face tracking scenarios (F1-F5), each subject has her/his pose changed all the time. This results in a certain challenge for all the trackers, as it becomes very difficult to predict the direction of the faces. For example, Fig. 7 illustrates the tracking results of sequence "H2" using the aforementioned approaches. It clearly shows that at frames 171 and 243 where the subject's head turns around, the three classical approaches "CPR", "CEPR" and "SPF" experience larger errors than the proposed "SomePF" algorithm. This is mainly due to the fact that "CPR", "CEPR" and "SPF" lose consistent observations of colour and motion cues and hence the particles in the trackers gradually drifts away. The motion trajectories generated by



(a)



(b)

Figure 6: Error statistics ϵ in the comparison of different algorithms CPF, CEPF, SPF, SomePF, IVT and MilTrack (left to right) for the eight testing sequences. Error bars indicate standard deviations.

these trackers show the consequence. In Fig. 8, the case is much different from Fig. 7. At frame 54, all the "CPR", "CEPR" and "SPR" trackers are distracted by the face of the third party. However, the proposed "SomePF" tracker correctly localises the object, as the detected motion energy feature has been used as a representative descriptor for image correspondence in this case.

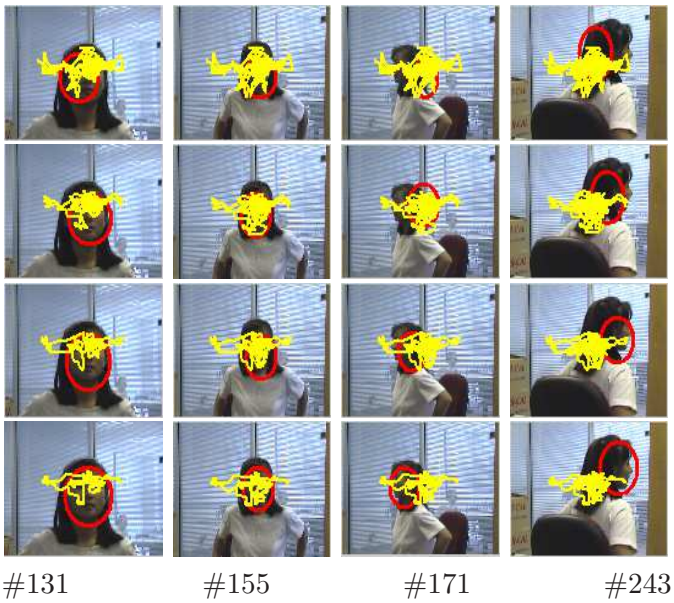


Figure 7: Examples of object tracking using different tracking algorithms for sequence "H2". Row 1 - "CPF", row 2 - "CEPF", row 3 - "SPF" and row 4 - "SomePF". Red circles indicate the tracks and yellow curves are the motion trajectories (and hereafter).

In the pedestrian scenarios P1-P3, human body is of flexible motion and therefore a tracker easily loses tracks due to self and/or mutual occlusions. In this subsection, we use two examples to demonstrate the performance of these trackers. Firstly, we look at the case shown in Fig. 9. In this scene, two persons repeatedly run across each other in the chamber. We expect to track the person who wears a red jacket. Due to the pose changes, the classical "CPR", "CEPR" and "SPF" trackers cannot obtain accurate locations of the target at frames 179 and 195. Comparably, the proposed "SomePF" scheme correctly locates the target because of the slow change occurring in the colour/motion energy features. Secondly, Fig. 10 illustrates the tracking results for a male pedestrian walking in a classroom. During the trial, the target hides himself behind a barrier. This experiment results in large scaling

Figure 8: Examples of object tracking using different tracking algorithms for sequence "H4".

and pose changes of the target. Similar to the previous pedestrian case, the proposed "SomePF" approach has better performance than the other three algorithms, as motion energy and colour features complement each other and overcome the track drifts.

4.3. Distractions and occlusions

In this subsection, we mainly discuss about the capabilities of different algorithms in handling distractions and occlusions. We start from the case of distractions due to neighbouring subjects. Here, our attention is to investigate color/edge/motion energy histograms and their characteristics using the aforementioned tracking algorithms.

As an example, Fig. 11 illustrates the intermediate results of different tracking algorithms. In this case, the third party gradually leaves the neighbourhood of the target subject. We introduce experimental results such as distributed particles, bar charts of edge and motion energy features of the facial areas. Fig. 11 denotes that (1) the classical "CPF" is reluctant to cope with the distraction problem as the tracks are attracted to the neighbouring faces with similar colors, and (2) both edge and motion energy features can be used as an effective tool to contend the distractions. These results can be

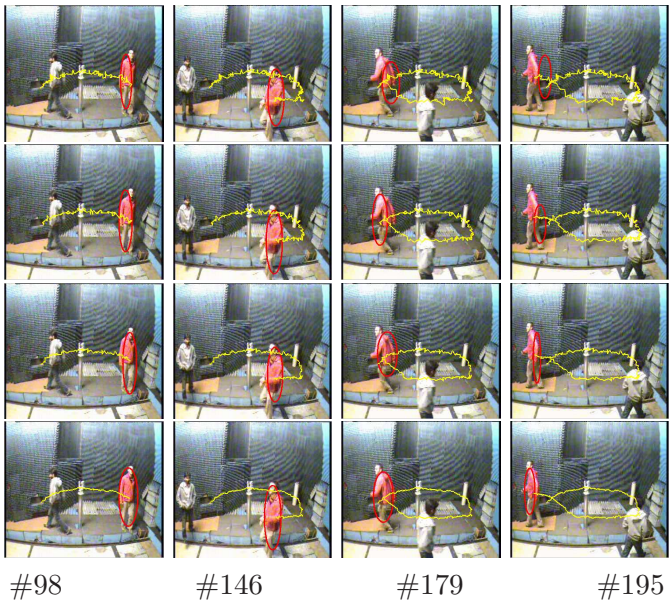


Figure 9: Examples of object tracking using different tracking algorithms for sequence "P1".

found on rows 2 and 3 of Fig. 11. They justify the consistent performance of the edge and motion energy based approaches.

In the case of occlusions, for the demonstrative purpose, we show color, edge and motion energy changes against time. Performance of the proposed tracking system, shown in Fig. 12, reveals that as another face moves into the scene, color and edge distributions accordingly change. We also observe that motion energy features are affected modestly (significant changes occur in the last image though).

4.4. Illumination changes and scaling

We here evaluate the performance of different tracking algorithms in illumination changes and image scaling. Each tracking system has to make significant efforts to accommodate lighting or size changes. In Fig. 13, we show an example out of video H5, where the subject continuously makes lateral, forward and backward moves that accompany various lighting conditions and scaled facial areas. Note that all the trackers are re-initialised at frame 309 for a fair comparison. These experimental results show that the proposed "SomePF" achieves the best tracking performance but the other

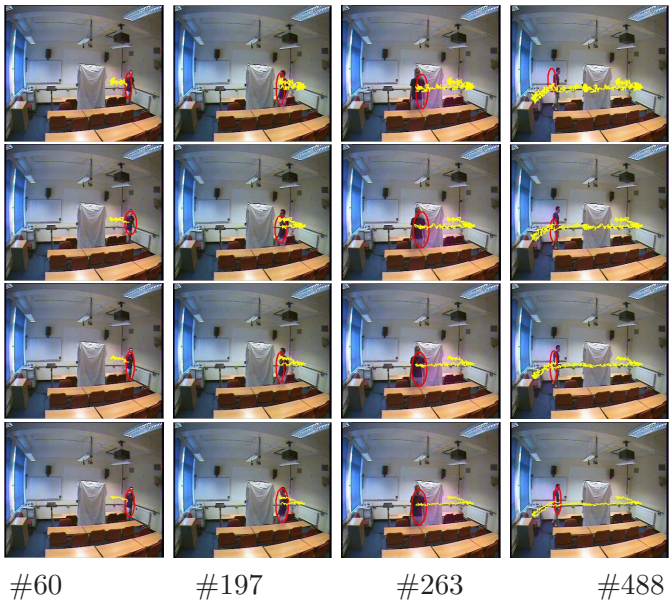


Figure 10: Examples of object tracking using different tracking algorithms for sequence "P3".

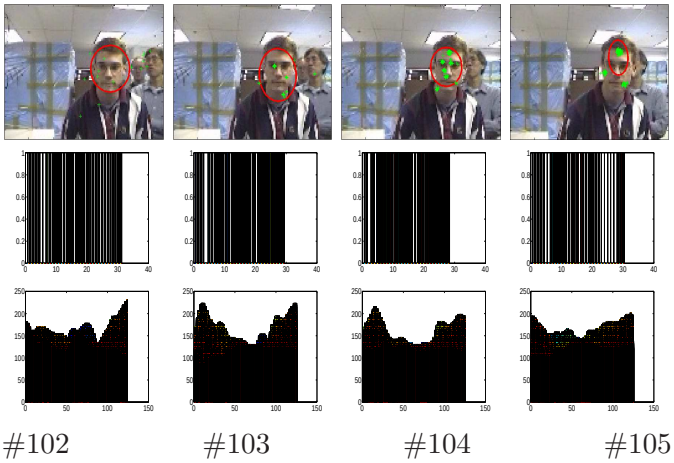


Figure 11: Distractions due to the subjects nearby: outcomes of several tracking algorithms for sequence "H4". Row 1 - "CPF" (red circles indicate the tracks whilst green additions are the centres of particles), row 2 - bar charts of edge features of the frontal face of the male subject (indicating the appearance change), and row 3 - bar charts of spatial-temporal motion energy features of the frontal face of the male subject.

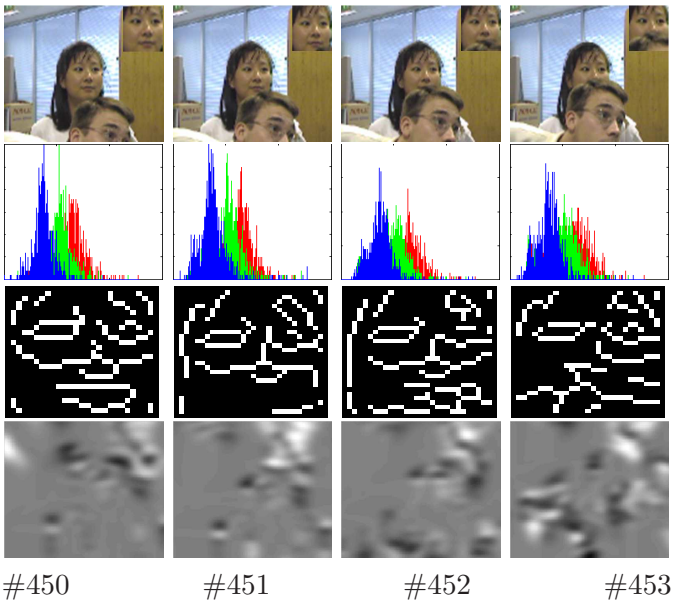


Figure 12: Occlusions: outcomes of our algorithm applied to the detected parts in sequence "H2". Row 1 - original images with the detected parts shown on the top right corner, row 2 - RGB distributions of the detected region, row 3 - Canny edge maps of the female's facial area, and row 4 - spatial-temporal motion energy features of the female's facial area. Better viewed in colour.

Figure 13: Illumination changes and scaling: outcomes of different tracking algorithms for sequence "H5". Row 1 - "CPF", row 2 - "CEPF", row 3 - "SPF" and row 4 - "SomePF".

classical methods gradually lose the target. For example, the "SPF" approach obtains the tracks located on the top of the target region. This is because the motion energy of the facial area is similar to that of the wall, which pushes the estimated location towards the upper boundary.

4.5. Comparison against "IVT" and "MilTrack"

"IVT" and "MilTrack" are two typical examples that were well developed in the last few years. Because of their elegant performance, these trackers have been frequently used in comparison against newly established tracking systems. In this subsection, we compare the proposed algorithm with "IVT" and "MilTrack". For example, Fig. 14 shows the errors (Euclidean distance between the tracks and the ground truth) of individual tracking algorithms for sequence "H1". Fig. 15 illustrates some image examples superimposed by the tracks using the proposed "SomePF" and classical "IVT" and "MilTrack" for the image sequence "H1". As observed, our "SomePF" algorithm has the capability of continuously tracking and quickly reacquiring the targets. Comparably, both "IVT" and "MilTrack" trackers lose the targets in images 17 and 34, respectively.

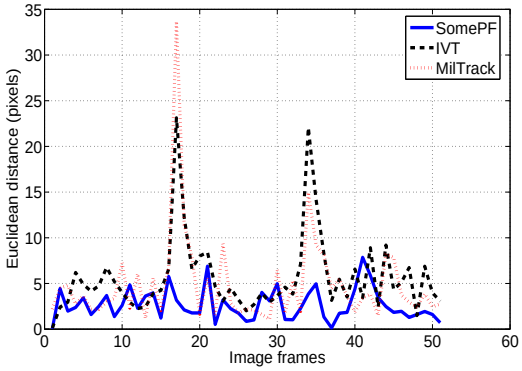


Figure 14: Euclidean distance between the tracks and the ground truth using different tracking algorithms for sequence "H1".

Figure 15: Comparisons of different tracking algorithms for sequence "H1". Red - "SomePF", green - "MilTrack", purple - "IVT". Better viewed in color.

4.6. Comparison against the trackers of [24] and [45]

In the subsection, we compare our method "SomePF" against two state of the art techniques, which have been reported in [24] and [45], respectively. Initialisations were set to these methods and the parameters of these methods were adjusted to obtain the best performance individually. To explore the the characteristics of each method, we use the test datasets created by Kwon and Lee [45], which consist of *skating*1, *soccer*, *ironman* and *matrix* with severe illumination, viewpoint changes, occlusions and motion blurs.

Table 2 shows the centre location errors in pixels of each algorithm for all the image sequences. This illustrates that the proposed tracking algorithm "SomePF" has achieved comparable or better performance than the other two methods in most of the image sequences. Our "SomePF" system employs the features that can be used to boost each other to handle different levels of noise, occlusions and illumination changes, where motion

#8 #26 #56 #69

Figure 16: Comparisons of different tracking algorithms for the challenging *matrix* sequence. Red - "SomePF", green - [24], white - [45]. Better viewed in color.

energy is used to complement the failure of color detection. The tracker of [24] achieves good performance in the *ironman* and *matrix* sequences but cannot maintain consistent tracking performance in the *skating*1 and *soccer* sequences, as its orientation estimations seem to be affected by the noise buried in the images. This tabulation also shows that the tracker of [45] has consistent performance throughout the entire dataset. As an example, Fig. 16 illustrates the tracking performance of different trackers for the *matrix* sequence, where the proposed "SomePF" tracker has better tracking performance than the others.

5. Conclusions and future work

In this paper, we have presented a particle filter based tracking algorithm that integrates measurements from colour and spatio-temporal motion energy components. In order to maintain the performance of a tracker in the presence of illumination changes and occlusions, we calculated the distribution of motion energy within each single particle. Since motion energy reflects the dynamic characteristics of a target, our tracker, using this feature, has achieved robustness to illumination changes and temporal occlusions. Afterwards, the colour and motion energy cues were adaptively fused for updating the observations. The proposed strategy has been compared

Table 2: Comparison of tracking accuracy of our tracker "SomePF", [24] and [45] against the overall image sequences. We run these algorithms three times and average the results.

Seq.	Image no.	[24]	[45]	"SomePF"
skating1	400	16	10	11
soccer	392	34	27	25
ironman	166	21	19	17
matrix	100	18	16	14

to five classical techniques, and the experimental results demonstrated that the proposed approach had better accuracy than the other state of the art techniques.

One of the major areas, which require our close attention in the next stage, is the capability of tracking small objects using the proposed algorithm. Up to date, we have observed that for indoors sequences, where the cameras are 4-5 metres (or less) away from the subjects, the proposed tracking algorithm seems to work well. In case the distance is larger than 8 metres, useful features can gradually fade out so that the correspondence across image frames become unstable. Another area for us to work on is to extend the case of tracking single subjects to that of matching multiple subjects, which appears more challenging in terms of uncertainty and ambiguity. In the future, we may investigate the integration of further global features which demonstrate robust invariance against extreme circumstances, where both colour and motion energy cues become unavailable.

6. Acknowledgment

The work of H. Zhou is supported in part by UK EPSRC (Grant EP/J006238/1) and Invest NI. X. Li is supported by the National Natural Science Foundation of China (Grant No: 61125106) and Shaanxi Key Innovation Team of Science and Technology (Grant No: 2012KCT-04).

References

[1] X. Liu, L. Lin, S. Yan, H. Jin, W. Jiang, Adaptive object tracking by learning hybrid template online, IEEE Trans. Circuits Syst. Video Techn. 21 (11) (2011) 1588–1599.

[2] N. Widynski, S. Dubuisson, I. Bloch, Integration of fuzzy spatial information in tracking based on particle filtering, IEEE Trans. Sys. Man Cyber. Part B 41 (3) (2011) 635–649.

[3] L. Li, S. Yan, X. Yu, Y. Tan, H. Li, Robust multiperson detection and tracking for mobile service and social robots, IEEE Transactions on Systems, Man, and Cybernetics, Part B 42 (5) (2012) 1398–1412.

[4] Z. Husz, A. Wallace, P. Green, Tracking with a hierarchical partitioned particle filter and movement modelling, IEEE Trans. Sys. Man Cyber. Part B 41 (6) (2011) 1571–1584.

[5] X. Zhang, W. Li, W. Hu, H. Lin, S. Maybank, Bloack covariance based l_1 tracker with a subtle template dictionary, Pattern Recognition 46 (7) (2013) 1750–1761.

[6] S. Zhang, H. Yao, X. Sun, X. Lu, Sparse coding based visual tracking: Review and experimentl comparison, Pattern Recognition 46 (7) (2013) 1772–1788.

[7] P. Cui, L.-F. Sun, F. Wang, S.-Q. Yang, Contextual mixture tracking, IEEE Trans. Multi. 11 (2009) 333–341.

[8] T. Yu, Y. Wu, Collaborative tracking of multiple targets, in: IEEE Conf. on Comput. Vis. and Patt. Recog., 2004, pp. 834–841.

[9] K. Hariharakrishnan, D. Schonfeld, Fast object tracking using adaptive block matching, IEEE Trans. Multi. 7 (5) (2005) 853–859.

[10] M. Lee, R. Nevatia, Human pose tracking in monocular sequence using multilevel structured models, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 27–38.

[11] D. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, International Journal of Computer Vision 77 (1-3) (2008) 125–141.

[12] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 983–990.

[13] Q. Wang, F. Chen, J. Yang, W. Xu, M.-H. Yang, Transferring visual prior for online object tracking, IEEE Transactions on Image Processing 21 (7) (2012) 3296–3305.

[14] C.-T. Hsu, M.-S. Hsieh, Region tracking for non-rigid video objects in a non-parametric map framework, Sig. Proc.: Image Comm. 21 (3) (2006) 235–251.

[15] L. Zhang, B. Wu, R. Nevatia, Detection and tracking of multiple humans with extensive pose articulation, in: Int'l Conf. on Comput. Vis., 2007, pp. 1–8.

[16] V. Reilly, H. Idrees, M. Shah, Detection and tracking of large number of targets in wide area surveillance, in: European conference on computer vision conference on Computer vision, 2010, pp. 186–199.

[17] H. Zhou, Y. Yuan, C. Shi, Object tracking using sift features and mean shift, Computer Vision and Image Understanding 113 (3) (2009) 345–352.

[18] H. Zhou, A. Wallace, P. Green, Efficient tracking and ego-motion recovery using gait analysis, Signal Processing 89 (12) (2009) 2367–2384.

[19] A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 798–805.

[20] K. Cannons, J. Gryn, R. Wildes, Visual tracking using a pixelwise spatiotemporal oriented energy representation, in: Proc. of the 11th European conference on Computer vision, 2010, pp. 511–524.

[21] N. Anjum, A. Cavallaro, Multifeature object trajectory clustering for video analysis, IEEE Trans. Circuits Syst. Video Techn. 18 (11) (2008) 1555–1564.

[22] H. Zhou, M. Taj, A. Cavallaro, Target detection and tracking with heterogeneous sensors, IEEE J. Selected Topics in Signal Process. 2 (4) (2008) 503–513.

[23] E. H. Adelson, J. R. Bergen, Spatiotemporal energy models for the perception of motion, J. Opt. Soc. Am. A 2 (2) (1985) 284–299.

[24] E. Maggio, F. Smerladi, A. Cavallaro, Adaptive multifeature tracking in a particle filtering framework, IEEE Trans. Circuits Syst. Video Techn. 17 (10) (2007) 1348–1359.

[25] D.-Y. Chen, K. Cannons, H.-R. Tyan, S.-W. Shih, H.-Y. M. Liao, Spatiotemporal motion analysis for the detection and classification of moving targets, IEEE Transactions on Multimedia 10 (8) (2008) 1578–1591.

[26] D. Tsai, M. Flagg, J. Rehg, Motion coherent tracking with multi-label mrf optimization, in: Proc. of the British Machine Vision Conference, 2010, pp. 56.1–56.11.

[27] G. Welch, G. Bishop, An introduction to the kalman filter, Tech. rep., University of North Carolina at Chapel Hill, NC, USA (1995).

[28] M. Isard, A. Blake, Condensation - conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1) (1998) 5–28.

[29] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 142–151.

[30] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proc. of International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.

[31] J. Shi, C. Tomasi, Good features to track, Tech. rep., Cornell University, Ithaca, NY, USA (1993).

[32] 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 Analysis and Machine Intelligence 18 (2) (1996) 138–150.

[33] T.-J. Cham, J. Rehg, A multiple hypothesis approach to figure tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 2239–2245.

[34] V. Pavlovic, J. Rehg, T.-J. Cham, K. Murphy, A dynamic bayesian network approach to figure tracking using learned dynamic models, in: IEEE International Conference on Computer Vision, 1999, p. 94.

[35] Y. Chen, Y. Rui, T. Huang, JPDAF based HMM or real-time contour tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 543–550.

[36] Y. Yoon, A. Kosaka, A. Kak, A new kalman-filter-based framework for fast and accurate visual tracking of rigid objects, IEEE Transactions on Robotics 24 (5) (2008) 1238–1251.

[37] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, P. J. Nordlund, Particle filters for positioning, navigation, and tracking, IEEE Trans. on Signal Processing 50 (2002) 425–437.

[38] M. Li, W. Chen, K. Huang, T. Tan, Visual tracking via incremental self-tuning particle filtering on the affine group, in: IEEE Conf. on Comput. Vis. and Patt. Recogn., 2010, pp. 1315–1322.

[39] J. Vermaak, N. Lawrence, P. Perez, Variational inference for visual tracking, in: IEEE Conf. Computer Vision and Pattern Recog, 2003, pp. 773–780.

[40] K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Trans. Info. Theo. 21 (1) (1975) 32–40.

[41] A. Yilmaz, Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.

[42] D. Freedman, P. Kisilev, Fast mean shift by compact density representation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1818–1825.

[43] Y.-J. Yeh, C.-T. Hsu, Online selection of tracking features using adaboost, IEEE Trans. Cir. and Sys. for Video Technol. 19 (2009) 442–446.

[44] N. M. Artner, A comparison of mean shift tracking methods, in: 12th Central European Seminar on Computer Graphics, 2008, p. 197C204.

[45] J. Kwon, K. Lee, Tracking by sampling trackers, in: ICCV, 2011, pp. 1195–1202.

[46] J. Kwon, K. Lee, Wang-landau monte carlo-based tracking methods for abrupt motions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (4) (2013) 1011–1024.

[47] M. Schaap, I. Smal, C. Metz, T. van Walsum, W. Niessen, Bayesian tracking of elongated structures in 3d images, in: 20th International Conference on Information Processing in Medical Imaging, 2007.

[48] J. Cui, S. Acton, Z. Lin, A monte carlo approach to rolling leukocyte tracking in vivo, Medical Image Analysis 10 (4) (2006) 598–610.

[49] J.-Z. Cheng, C.-M. Chen, E. Cole, E. Pisano, D. Shen, Automated delineation of calcified vessels in mammography by tracking with uncertainty and graphical linking techniques, IEEE Trans. Med. Imaging 31 (11) (2012) 2143–2155.

[50] A. Smith, T. Ledgeway, Motion detection in human vision: a unifying approach based on energy and features, Proceedings of the Royal Society of London: Series B 268 (2001) 1889–1899.

[51] Y. M. Lam, B. Shi, Extending position/phase-shift tuning to motion energy neurons improves velocity discrimination, in: J. Platt, D. Koller,

Y. Singer, S. Roweis (Eds.), Advances in Neural Information Processing Systems 20, MIT Press, Cambridge, MA, 2008, pp. 809–816.

[52] Q. Zaidi, J. S. DeBonet, Motion energy versus position tracking: spatial, temporal, and chromatic parameters, Vision Research 40 (2000) 3613–3635.

[53] K. Cannons, R. Wildes, Spatiotemporal oriented energy features for visual tracking, in: Proc. of Asian conference on Computer vision, 2007, pp. 532–543.

[54] C. Shan, T. Tan, Y. Wei, Real-time hand tracking using a mean shift embedded particle filter, Pattern Recogn. 40 (2007) 1958–1970.

[55] T. Xiang, S. Gong, Beyond tracking: Modelling activity and understanding behaviour, International Journal of Computer Vision 67 (1) (2006) 21–51.

[56] P. Cui, Z.-Q. Liu, L.-F. Sun, S.-Q. Yang, Hierarchical visual event pattern mining and its applications, Data Mining and Knowledge Discovery 22 (2011) 467–492.

[57] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.

[58] K. Nummiaro, E. Koller-Meier, L. V. Gool, An adaptive color-based particle filter., Image Vision Comput. (2003) 99–110.

[59] T. Kailath, The divergence and bhattacharyya distance measures in signal selection, IEEE Trans. Communication Technology 15 (1) (1967) 52–60.

[60] Z. Lu, G. Sperling, Three-systems theory of human visual motion perception: review and update, J. Opt. Soc. Am. A 18 (9) (2001) 2231–2270.

[61] Q. Hu, J. D. Victor, A set of high-order spatiotemporal stimuli that elicit motion and reverse-phi percepts, J. Vis. 10 (3) (2010) 1–16.

[62] C. Zitnick, N. Jojic, S. Kang, Consistent segmentation for optical flow estimation, in: Proc. of Int'l Conf. on Comput. Vis., 2005, pp. II: 1308–1315.

[63] M. Spengler, B. Schiele, Towards robust multi-cue integration for visual tracking, Lecture Notes in Computer Science 2095 (2001) 93–106.

[64] R. Jacobs, What determines visual cue reliability?, Trends in Cognitive Sciences 6 (8) (2002) 345–350.

[65] P. Ragothaman, T. Yang, W. Mikhael, R. Muise, A. Mahalanobis, Efficient adaptive subspace tracking algorithm for automatic target recognition, Electronics Letters 42 (2006) 1183–1184.

[66] J. Triesch, C. Von Der Malsburg, Democratic integration: Self-organized integration of adaptive cues, Neural Comput. 13 (2001) 2049–2074.

[67] B. Babenko, M.-H. Yang, S. Belongie, Robust object tracking with online multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1619–1632.