International Journal of Natural Computing Research (IJNCR), IGI Global, Vol. 4, No. 3, pp. 69-84, 2014.07-09

Visual tracking using multimodal particle filter

Tony Tung
Kyoto University, Japan

Takashi Matsuyama
Kyoto University, Japan

ABSTRACT

Visual tracking of humans or objects in motion is a challenging problem when observed data undergo appearance changes (e.g., due to illumination variations, occlusion, cluttered background, etc.). Moreover, tracking systems are usually initialized with predefined target templates, or trained beforehand using known datasets. Hence, they are not always efficient to detect and track objects whose appearance changes over time. In this paper, we propose a multimodal framework based on particle filtering for visual tracking of objects under challenging conditions (e.g., tracking various human body parts from multiple views). Particularly, we integrate various cues such as color, motion and depth in a global formulation. The Earth Mover distance is used to compare color models in a global fashion, and constraints on motion flow features prevent common drifting effects due to error propagation. In addition, the model features an online mechanism that adaptively updates a subspace of multimodal templates to cope with appearance changes. Furthermore, the proposed model is integrated in a practical detection and tracking process, and multiple instances can run in real-time. Experimental results are obtained on challenging real-world videos with poorly textured models and arbitrary non-linear motions.

1. INTRODUCTION

Visual tracking of human body parts is widely used in many real-world applications, such as video surveillance, games, cultural and medical applications (e.g., for motion and behavior study). The literature has provided successful algorithms to detect and track objects of a predefined class in image streams or videos (Yilmaz, Javed, & Shah, 2006; Wu, Lim, & Yang, 2013). Simple objects can be detected and tracked using various image features such as color regions, edges, contours, or texture. On the other hand, complex objects such as human faces require more sophisticated features to handle the multiple possible instances of the object class. For this purpose, statistical methods are a good alternative. First, a statistical model (or classifier) learns different patterns related to the object of interest (e.g., different views of human faces), including good and bad samples. And then the system is able to estimate whether a region contains an

object of interest or not. This kind of approach has become very popular. For example, the face detector of (Viola, & Jones, 2001) is well known for its efficiency. The main drawback is the dependence to prior knowledge on the object class. As the system is trained on a finite dataset, the detection is somehow constrained to it. As a matter of fact, most of the tracking methods were not designed to keep the track of an object whose appearance could strongly change. If there is no a priori knowledge on its multiple possible appearances, then the detection fails and the track is lost. Hence, tracking a head which turns completely, or tracking a hand in action remain challenging problems, as appearance changes occur quite frequently for human body parts in motion .

In order to leverage visual tracking under challenging conditions, we introduce a multimodal framework based on the well-known particle filter model (Isard, & Blake, 1998). Our global model integrates various cues such as color, motion, and also depth to perform robust tracking. In addition, Earth Mover distance (Rubner, Tomasi, & Guibas, 1998) has been chosen to compare color models due to its robustness to small color variations, and drift effects inherent to adaptive tracking methods are handled using extracted motion features (e.g., optical flows). As well, an online adaptive process updates a subspace of multimodal templates so that the tracking system remains robust to occlusions and appearance changes. The tracking system is integrated in a practical workflow containing two modes, switching between detection and tracking. The detection steps involve trained classifiers to update estimated positions of the tracking windows. In our experiments, we use the cascade of boosted classifiers of Haar-like features by (Viola, & Jones, 2001) to perform head detection. Other body parts can be either detected using this technique with ad-hoc training samples, or chosen by users at the initialization step (i.e., pick and track method), or as well can be deduced based on prior knowledge on human shape features and constraints. Our experimental results show accuracy and robustness of the proposed method on challenging video sequences of humans in motion. For example, we use videos of yoga performances (stretching exercises at various speeds) with poorly textured regions, and arbitrary non-linear motions were used for testing (see Fig. 1), and also multiple view videos of multiple people interacting during a group discussion in various environments (e.g., meeting room, conference hall) as can be seen in Sect. 5.

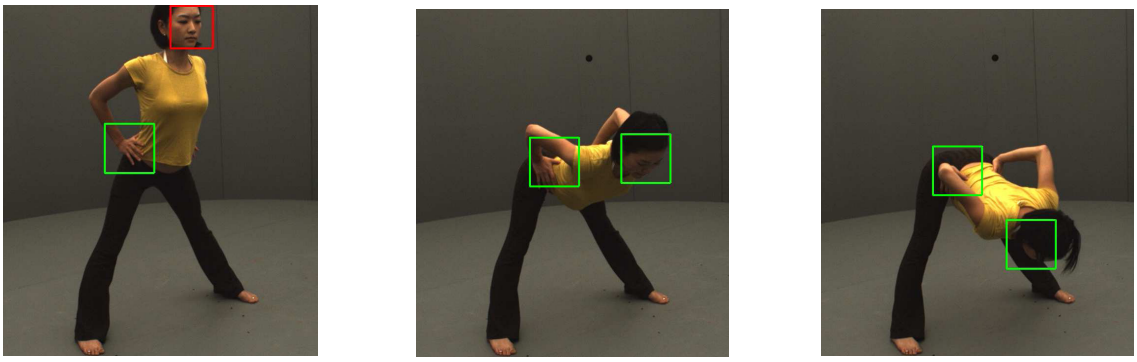


Fig. 1. Body part tracking with multimodal particle filter (using color and motion). Here, body parts located by the tracker are highlighted in green, while regions located by the detector (e.g., face) are highlighted in red. The proposed model is robust to strong appearance changes.

The rest of the paper is organized as follows. The next section gives a recap of work related to the techniques presented in this work. Section 3 presents an overview of the algorithm (initialization step and workflow). Section 4 describes the proposed multimodal particle filter framework. Section 5 presents experimental results on real-world datasets. Section 6 concludes with a discussion on our contributions.

2. RELATED WORK

During the past decades, image sensing devices such as video cameras and depth sensors have quickly become more accurate and accessible for non-expert users. This has lead to a rapid growth of various imaging applications (Tung, & Matsuyama, 2012). In particular, the scientific community has shown a real interest to human body part detection and tracking. For example, face detection in images is nowadays a popular and well explored topic (Viola, & Jones, 2001; Hjelmas, & Low, 2002; Choudhury, Schmid, & Mikolajczyk, 2003). In (Viola, & Jones, 2001), the authors proposed a cascade of boosted tree classifiers of Haar-like features. The classifier is first trained on positive and negative samples, and then the detection is performed by sliding a search window through candidate images and checking whether a region contains an object of interest or not. The technique is known to be fast and efficient, and can be tuned to detect any kind of object class if the classifier is trained on good samples.

Similarly, tracking in video is also a popular field of research as image streams are now ubiquitous. However, object recognition in video data is still challenging when resolution is low (e.g., in video surveillance) and noise is high (e.g., due motion blur). Various approaches were proposed to extract image features for pattern matching and tracking (Lucas, & Kanade, 1981; Tomasi, & Kanade, 1991; Lowe, 2004; Lucena, Fuertes, & de la Blanca, 2004; Tola, Lepetit, & Fua, 2008). Lucas, Tomasi and Kanade first select the good features which are optimal for tracking, and then keep the tracks of these features in consecutive frames. The KLT feature tracker is often used for optical flow estimation to estimate the deformations between two frames. As a differential method, it assumes that the pixel intensity of objects is not significantly different between two frames.

Techniques based on prediction and correction such as Meanshift (Cheng, 1995; Comaniciu, Ramesh, & Meer, 2000; Comaniciu, Ramesh, & Meer, 2003), Kalman filter (Kalman, 1960; Terzopoulos, & Szeliski, 1992; Blake, Curwen, & Zisserman, 1993; Rehg and Kanade, 1994), and more recently particle filters have become widely used (Isard, & Blake, 1998; Doucet, Godsill, & Andrieu, 2000; Perez, Hue, Vermaak, & Gangnet, 2002; Sugimoto, Yachi, & Matsuyama, 2003; Okuma, Taleghani, de Freitas, Kakade, Little, & Lowe, 2004; Dornaika, & Davoine, 2005; Wang, Chen, & Gao, 2005; Li, Ai, Yamashita, Lao, & Kawade, 2007; Ross, Lim, Lin, & Yang, 2007; Kim, Kumar, Pavlovic, & Rowley, 2008). Particle filters (or sequential Monte Carlo or Condensation) are Bayesian model estimation techniques based on simulation. The basic idea is to approximate a sequence of probability distributions using a large set of random samples (called particles). Then the particles are propagated through the frames based on importance sampling and resampling mechanisms. Usually, the particles converge rapidly to the distributions of interest. The algorithm allows robust tracking of objects in cluttered scene, and can handle non-linear motion models more complex than those commonly used in Kalman filters. The major differences between the different particle filter based approaches rely on the design of

the sampling strategies, which make particles having higher probability mass in regions of interest.

In (Black, & Jepson, 1998; Collins, Liu, & Leordeanu, 2005 ; Wang, Chen, & Gao, 2005; Ross, Lim, Lin, & Yang, 2007; Kim, Kumar, Pavlovic, & Rowley, 2008), linear dimension reduction methods (PCA, LDA) are used to extract feature vectors from the regions of interest. These approaches suit well for adaptative face tracking and can be formulated in the particle filtering framework as well. Nevertheless they require a big training data set to be efficient (Martinez, & Kak, 2001), and still cannot cope with unpredicted change of appearance. On the other hand, color-based models of regions can capture larger appearance variations (Bradski, 1998; Comaniciu, Ramesh, & Meeh, 2000). In (Perez, Hue, Vermaak, & Gangnet, 2002), the authors integrate a color-based model tracker (as in the Meanshift technique of Comaniciu, Ramesh, and Meeh) within a particle filter framework. The model uses color histograms in the HSV space and the Bhattacharyya distance for color distribution comparisons. Nevertheless these methods usually fail to track objects in motion or have an increasing drift on long video sequences due to strong appearance changes or important lighting variations (Matthews, Ishikawa, & Baker, 2004). Indeed most algorithms assume that the model of the target object does not change significantly over time. To adapt the model to appearance changes and lighting variations, subspace of the target object features are extracted (Collins, Liu, & Leordeanu, 2005; Wang, Chen, & Gao, 2005; Ross, Lim, Lin, & Yang, 2007; Kim, Kumar, Pavlovic, & Rowley, 2008). In (Ross, Lim, Lin, & Yang, 2007), a subspace of eigenvectors representing the target object is incrementally updated through the tracking process. Thus, offline learning step is not required and tracking of unknown objects is possible. Recently, (Kim, Kumar, Pavlovic, & Rowley, 2008) proposed to extend this approach with additional terms in the data likelihood definition. In particular, the drift error is handled using an additional dataset of images. However, these approaches are particularly tuned for face tracking, and still require training datasets for every different view of faces. Note that tracking of multiple similar objects using various constraints (e.g., contextual constraint learning, constant velocity assumption, etc.) are out-of-scope of this paper (Brendel, Amer, & Todorovic, 2011; Butt, & Collins, 2013).

The overall workflow of our approach divides into two steps which are detection and tracking, as (Sugimoto, Yachi, & Matsuyama, 2003; Li, Ai, Yamashita, Lao, & Kawade, 2007). Switching between the two modes allows dynamic updates of the search window to an accurate position whenever the detection is positive. In this work, we propose a multimodal particle filter which relies on various cues (e.g., color, motion, depth) to achieve robust visual tracking (see also Maggio, Smeraldi, & Cavallaro, 2007; Wang, & Tang, 2010). Our tracker uses a subspace of multimodal templates of regions of interest extracted from previous frames, and relies on them to estimate the position of the object in the current frame. The subspace is iteratively updated through the video sequence, and dynamically updated by the detection process. The detection is performed by a cascade of boosted classifiers (Viola, & Jones, 2001) and thus can be trained to detect any object class. We also propose to use the Earth Mover distance to improve the robustness of tracking with lighting variations (see also Karavasilis, Nikou, & Likas, 2011), and constraints based on optical flow estimations to cope with drift effects. Note that a preliminary version of this work was presented in (Tung, & Matsuyama, 2008).

3. TRACKING SYSTEM OVERVIEW

This section describes the overall workflow of the proposed visual tracking system, which consists of a practical tracking-by-detection strategy for object tracking in video. It contains two modes, switching between a detector and a multimodal tracker. The tracking process runs independently using multimodal cues (see Sect. 4) when no detection is positive for the class of the object of interest. It returns the object estimated position in the video (e.g., in pixel coordinates). On the other hand, the detector is used at system initialization, and to dynamically update the object estimated position.

3.1. Initialization

The initialization step consists in defining the objects to track. Basically, there are three practical strategies to define regions of interest (i.e., targets to track) as described below. Note that in this paper, we use human body parts to illustrate the effectiveness of our model because of the wide rang of possible applications.

1. Automatically using a trained detector or template matching, e.g., using statistical machine learning method such as the cascade classifier of Viola and Jones, or some geometrical structure assuming the object shape is known a priori (Tung, & Matsuyama, 2012).

2. Manually by interactively picking regions of interest (i.e., initialization in first frame). This allows users to track various body parts regardless of any prior knowledge.

3. Incidentally using a priori knowledge (e.g., fuzzy rules) and by reasoning on the object constraints. As constrained motions of the human body (assuming its structure is known) give cues on body part locations, e.g., face position gives hints to deduce torso position, etc.

After initialization, a subspace of reference templates consisting of the regions of interest (and containing multimodal information) is used by the tracker when processing the following frames. In some of the experimental results presented in Sect. 5, we have combined the three approaches: the head of a subject is automatically detected using a face detector, then torso position is deduced (assuming initial standing position), while hand positions are chosen manually in the first frame by picking.

3.2. Workflow

Assuming initialization occurs at time t_0, then for every frame at $t, t > t_0$, the tracker estimates the positions of M objects of interest $\{A_i\}_{i=1...M}$ based on a multimodal template-model subspace $S_t^i = \{h_{t-k}^i, ..., h_{t-1}^i\}$ as introduced above, where h_j^i denotes the color-model of A_i at time j, and k is the size of the subspaces (which in fact can be different for every object). Assuming a Bayesian framework (see Sect. 4), the state x_t^i corresponding to the estimated position of A_i at time t by the tracker, is inferred by S_t^i and x_{t-1}^i. We denote by y_t^i the output corresponding to the detection of A_i at time t, and z_t^i the output of the overall system. If the

detection of A_i at t is positive, then $z_t^i = y_t^i$, else $z_t^i = x_t^i$. Thus, if the detection of A_i at t is positive, then S_{t+1}^i is updated using the color model corresponding to y_t^i. And if not, then S_{t+1}^i is updated using the color model corresponding to x_t^i. The workflow is illustrated on Figure 2 with $M = 1$ and $k = 1$.

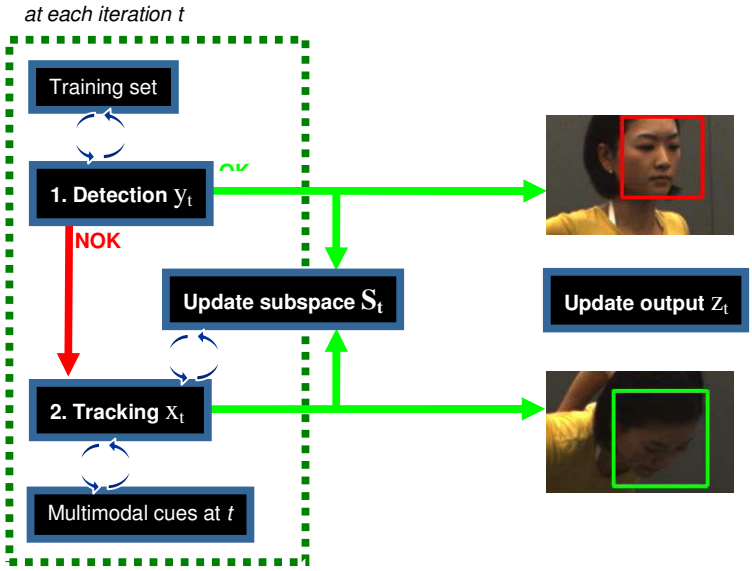


Fig. 2. Overall workflow of the multimodal tracking system. If the detection process y_t at time t is positive, then the system output z_t and the subspace of multimodal templates S_t are updated using y_t. If the detection fails, then z_t and S_t are updated using the tracking output x_t. Note that S_t as well as multimodal cues are used to estimate x_t.

4. MULTIMODAL PARTICLE FILTER

In this section we present our particle filter-based multimodal framework. The global formulation can take into account various multiple cue, such as a color-based model (Isard, & Blake, 1998; Perez, Hue, Vermaak, & Gangnet, 2002), motion estimation by optical flows (Tomasi, & Kanade, 1991), and depth information. The Earth Mover Distance (Rubner, Tomasi, & Guibas, 1998) is used to compute distances between color models while being robust to lighting variations. Motion features and depth information are also used to improve tracking accuracy. Moreover, as described above, our method updates iteratively a subspace of multimodal templates to handle appearance changes and partial occlusions.

4.1. Particle filter-based multimodal framework

We denote by x_t a target state at time t, z_t the observation data at time t, and $Z_t = \{z_1,..., z_t\}$ all the observations up to time t. Assuming a non-Gaussian state space model, the prior probability $p(x_t \mid Z_{t-1})$ at time t in a Markov process is defined as:

$$p(x_t \mid Z_{t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Z_{t-1}) dx_{t-1}, \tag{1}$$

where $p(x_t \mid x_{t-1})$ is a state transition distribution, and $p(x_{t-1} \mid Z_{t-1})$ stands for a posterior probability at time $t-1$. The posterior probability whose the tracking system aims to estimate at each time is defined as:

$$p(x_t \mid Z_t) \propto p(z_t \mid x_t) p(x_t \mid Z_{t-1}), \tag{2}$$

where $p(z_t \mid x_t)$ is the data likelihood at time t. According to the particle filtering framework, the posterior $p(x_t \mid Z_t)$ is approximated by a Dirac measure on a finite set of P particles $\{x_t^i\}_{i=1...P}$ following a sequential Monte Carlo framework (Doucet, Godsill, & Andrieu, 2000). Candidate particles are sampled by a proposal transition kernel $q(\tilde{x}_t^i \mid x_{t-1}^i, z_{t-1})$. The new filtering distribution is approximated by a new sample set of particles $\{\tilde{x}_t^i\}_{i=1...P}$ having the importance weights $\{w_t^i\}_{i=1...P}$, where

$$w_t^i \propto \frac{p(z_t \mid \tilde{x}_t^i) p(\tilde{x}_t^i \mid x_{t-1}^i)}{q(\tilde{x}_t^i \mid x_{t-1}^i, z_{t-1})} \qquad \textit{and} \qquad \sum_{i=1}^{P} w_t^i = 1. \tag{3}$$

The sample set $\{x_t^i\}_{i=1...P}$ can then be obtained by resampling $\{\tilde{x}_t^i\}_{i=1...P}$ with respect to $\{w_t^i\}_{i=1...P}$. By default, the Bootstrap filter is chosen as proposal distribution: $q(\tilde{x}_t^i \mid x_{t-1}^i, z_{t-1}) = p(\tilde{x}_t^i \mid x_{t-1}^i)$. Hence the weights can be computed by evaluating the corresponding data likelihood. Finally, x_t is estimated upon the Monte Carlo approximation of the expectation $\hat{x}_t = \frac{1}{P}\sum_{i=1}^{P} x_t^i$.

We denote by E, a global (multimodal) energy function that can be defined using multiple various cues. For example, in our experiments: $E = E_s + E_m + E_d + E_z$, where E_s is an energy related to color cues (see Sect. 4.2), E_m and E_d are energies related to motion features (see Sect. 4.3), and E_z is an energy related to depth information when available (see Sect. 4.4). E has lower values as the search window is close to the target object. Thus, to favor candidate regions (i.e., samples) whose multimodal information are similar to the reference model at time t, the data likelihood $p(z_t \mid x_t)$ is modeled as a Gaussian function:

$$p(z_t \mid \tilde{x}_t^i) \propto \exp\left(-\frac{E}{\sigma^2}\right), \tag{4}$$

where σ is a scale factor, and therefore a small E returns a large weight.

4.2. Color cue

4.2.1. Color-based model

The efficiency of color distributions to track color content of regions that match a reference color model has been demonstrated in (Bradski, 2000; Comaniciu, Ramesh, & Meeh, 2000; Perez, Hue, Vermaak, & Gangnet, 2002). They are represented by histograms to characterize the chromatic information of regions. Hence they are robust against non-rigidity and rotation. In addition, the Hue-Saturation-Value (HSV) color space has been chosen due to its low sensitivity to lighting condition. In our approach, color distributions are discretized into three histograms of N_h, N_s, and N_v bins for the hue, saturation, and value respectively.

Let α be h, s, or v, $q_t(x_t) = \frac{1}{3}\sum_{\alpha} q_t^{\alpha}(x_t)$, and $q_t^{\alpha}(x_t) = \{q_t^{\alpha}(i, x_t)\}_{i=1...N_{\alpha}}$. $q_t(x_t)$ denotes the kernel density estimate of the color distribution in the candidate region $R(x_t)$ of the state x_t at time t, and is composed by:

$$q_t^{\alpha}(i, x_t) = K_{\alpha} \sum_{u \in R(x_t)} \delta[h_{\alpha}(u) - i], \tag{5}$$

where K_{α} is a normalization constant so that $\sum_{i=1}^{N_{\alpha}} q_t^{\alpha}(i, x_t) = 1$, h_{α} is a function assigning the pixel color at location u to the corresponding histogram bin, and δ is the Kronecker delta function.

At time t, $q_t(x_t)$ is compared to a set of reference color model templates $S_t = \{h_{t-k}, ..., h_{t-1}\}$, where k is the number of templates. The templates are extracted iteratively from the detected regions at each frame. We recall that color model subspaces help to handle appearance changes and partial occlusions, and we define the energy function:

$$E_s[S_t, q_t(x_t)] = \min_{h \in S_t} (D^2[h, q_t(x_t)]), \tag{6}$$

where D is a distance between color distributions (see Sect. 4.2.2).

4.2.2 Earth Mover distance

We propose to use the Earth Mover distance (EMD) (Hillier, & Lieberman, 1990; Rubner, Tomasi, & Guibas, 1998) to strengthen the property of invariance to lighting of the HSV color space. EMD allows the global comparison of color distributions relying on a global optimization process. This method is more robust than approaches relying on histogram bin-to-bin distances that are more sensitive to quantization and small color changes. The distributions are represented by sets of weighted features called *signatures*. The EMD is then defined as the minimal amount of *work* needed to match a signature to another one. The notion of work relies on a metric (e.g. a distance) between two features. In our framework we use the L_1 norm as distance, and histogram bins as features.

Assuming two signatures to compare $P = \{(p_1, w_1),...,(p_m, w_m)\}$ and $Q = \{(q_1, u_1),...,(q_n, u_n)\}$, P having m components p_i with weight w_i, and Q having n components q_j with weight u_j. The global optimization process consists in finding the amount of data f_{ij} of a signature to be transported from the component i to the component j that minimizes the work W:

$$W = \min_{f_{ij}} (\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}), \tag{7}$$

where d_{ij} is the distance between the components p_i and q_j assuming the following constraints:

$$\begin{aligned}
&f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq n,\\
&\sum_{j=1}^{n} f_{ij} \leq w_i \qquad 1 \leq i \leq m,\\
&\sum_{i=1}^{m} f_{ij} \leq u_j \qquad 1 \leq j \leq n,\\
&\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_i, \sum_{j=1}^{n} u_j).
\end{aligned}$$

The first constraint allows only the displacements from P to Q. The two following constraints bound the amount of data transported by P, and the amount of data received by Q to their respective weights. The last constraint sets the maximal amount of data that can be displaced.
The EMD distance D between two signatures P and Q is then defined as:

$$D(P,Q) = \frac{W}{\mathrm{N}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}, \tag{8}$$

where the normalization factor N ensures a good balance when comparing signatures of different size (N is the smallest sum of the signature weights). Note that EMD computation can be approximated in linear time with guaranteed error bounds (Shirdhonkar, & Jacobs, 2008).

4.3 Motion cues

Tracking using color information alone, as seen in many conventional tracking systems, usually leads to error propagation over time and drift (Matthews, Ishikawa, & Baker, 2004). Actually, structural information of templates is lost when evaluating sample positions as only the mass of colors is taken into account. Hence, we propose to use motion features to guide the search window through the tracking process, as they are based on local features (e.g., corners). In our implementation, motion features are extracted using the KLT feature tracker (Lucas, & Kanade,

1981; Tomasi, & Kanade, 1991), although other techniques can be also applied (Lowe, 2004; Lucena, Fuertes, & de la Blanca, 2004; Tola, Lepetit, & Fua, 2008). The method detects local features and matches similar ones between consecutive frames (see Fig. 3). Outliers are filtered by RANSAC.

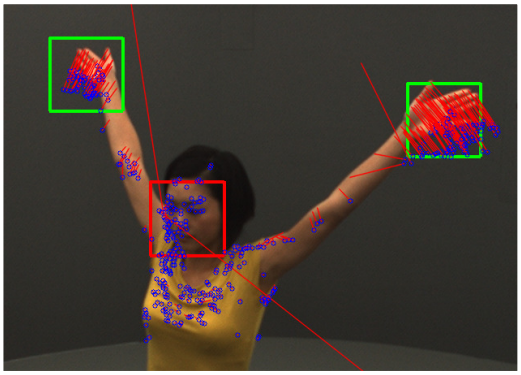


Fig. 3. Motion features. Motion features are extracted to support tracking process. Blue dots denote feature positions in the previous frame. Red lines show the estimated motion flows. Note that outliers are filtered by RANSAC.

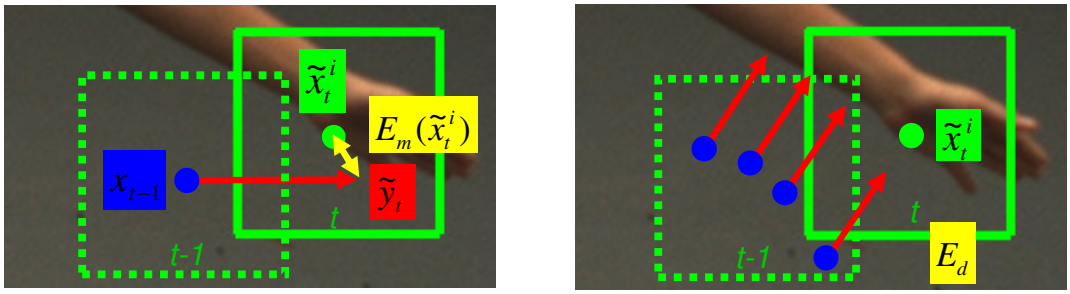
Assuming the set $Y_{t-1} = \{y_{t-1}^j\}_{j=1...m}$ of m motion features detected in the neighborhood region of the state x_{t-1} (see Sect. 4) at time $t-1$, and the set $Y_t = \{y_t^j\}_{j=1...m}$ of matching features extracted at time t, then $(Y_{t-1}, Y_t) = \{(y_{t-1}^j, y_t^j)\}_{j=1...m}$ forms a set of m motion vectors (optical flow field) between the frames at time $t-1$ and t. As well, we denote by $\tilde{Y}_t^i$ the set of features detected in the neighborhood region of the particle $\tilde{x}_t^i$, and $\tilde{y}_t$ the position of the search window estimated by optical flow as: $\tilde{y}_t = x_{t-1} + \text{median}(\{y_{t-1}^j - y_t^j\}_{j=1...m})$. Thus we define the following energy functions:

$$E_m(\tilde{x}_t^i) = \alpha \cdot \left\| \tilde{x}_t^i - \tilde{y}_t \right\|_2 \qquad \textit{and} \qquad E_d = \beta \cdot \mathrm{C}(\tilde{Y}_t^i, Y_t), \tag{9}$$

where α and β are two constant values, and C is the following function:

$$\mathrm{C}(\tilde{Y}_t^i, Y_t) = 1 - \frac{\text{card}(\tilde{Y}_t^i \cap Y_t)}{\text{card}(Y_t)}. \tag{10}$$

The data energy E_m aims to favor the particles located around the object target position estimated by optical flow, whereas E_d aims to prevent the drift effect. E_d works as a constraint which attracts the particles near the estimated search window (see Fig. 4). E_m and E_d are introduced in the overall energy formulation as described in Sect. 4.1.



(a) (b)

Fig. 4. Motion cues. (a) E_m measures the distance between the estimated position $\tilde{x}_t^i$ by particles and the estimated position by optical flow $\tilde{y}_t$. (b) E_d maximizes the number of features detected in the previous frame.

4.4 Depth cues

The multimodal framework allows the introduction of various different cues. Particularly, in some contexts where multiple video cameras are used (and geometrically calibrated), such as in 3D video studios (Matsuyama, Nobuhara, Takai, & Tung, 2012) or in smart environments (Tung, Gomez, Kawahara, & Matsuyama, 2012), it is possible to retrieve depth information from each camera view point using stereo reconstruction methods, and even obtain 3D reconstruction of the observed scene (Hartley, & Zisserman, 2004). As well, it is possible to get depth information using depth sensors, such as time-of-fight cameras or structured light cameras (e.g., Microsoft Kinect), and align geometrically depth images to color images obtained using conventional video cameras. Hence, depth cues can be introduced in the global energy E (see Sect. 4.1) by defining an energy function E_z as follows:

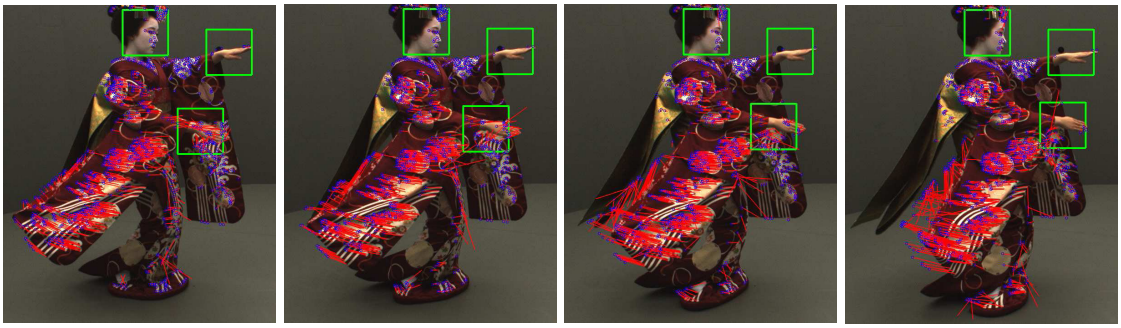
$$E_z[S_t, q_t(z_t)] = \min_{h_z \in S_t} (D_z^2[h_z, q_t(z_t)]), \tag{11}$$

where, z_t is the depth value corresponding to color image pixel at x_t, $q_t(z_t)$ denotes the kernel density estimate of depth distribution defined as in Eq. 5, D_z is a distance between depth distribution, and h_z is an element of the multimodal subset S_t containing depth information (see Eq. 6).

5. EXPERIMENTAL RESULTS

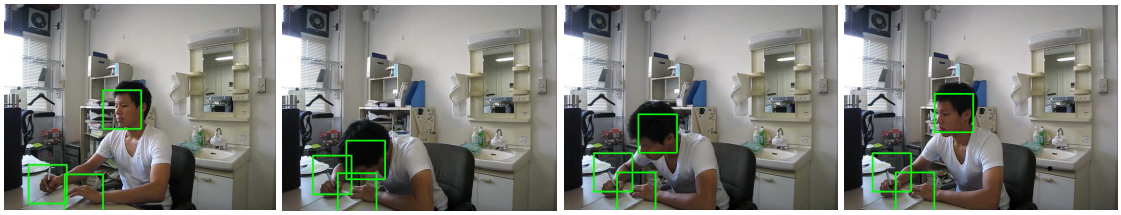
Our algorithm has been tested on numerous real-world video sequences, and using different cue combinations. First, we have tracked the body parts of a lady practicing yoga (head, hands, torso, and feet) in different video sequences and from different viewpoints using color and motion cues. The model wears simple clothes with no additional features (see Fig. 1 and Fig. 7). As well, we have tested the same tracker on traditional Japanese dancer wearing clothes which are more much complex and contain a lot of features (see Fig. 5). The video resolutions are 640x480 and 720x576 pixels respectively and were acquired at 25 fps. The algorithm was run on a Core2Duo 3.0 GHz with 4GB RAM. As observed, different body parts are successfully tracked simultaneously. Furthermore, the tracking system was tested on multiple subjects captured from multiple viewpoints using color and motion cues in several long sequences of about 11min at 25fps (see samples in Fig. 8a, 8b, 8c), and also using additional depth cues in several sequences

of 3min and 15min at 25fps (see samples in Fig. 8d, 8e). We computed head poses by fitting a 3D face model. Face orientations using pitch, roll and yaw angles are estimated with 5 deg of accuracy. Roll angles can be estimated in a range of at least [-70;70], which is already significantly larger compared to state-of-the-art system implementation such as Kinect SDK. The system was partially implemented on the GPU, and the tracking of 4 different targets was achieved in realtime. The following parameters were identical for all the experiments: we have used $N_h = 10$, $N_s = 10$ and $N_v = 10$ for the quantization of color models, $P = 200$ particles, $k = 5$ for the color model subspace size, and $\sigma^2 = 0.1$ as scale factor of the likelihood model. The constant values α and β weight the contribution of the motion cues, and are tuned regarding to the frame size. He have defined a square window size of 40 pixels to determine the regions of interest. The proposed formulation has shown promising results even in uncontrolled environments. The Figures 1 and 6 illustrate the robustness to appearance change, lighting variation and partial occlusion, thanks to the online update of the color-based model subspace combined with the Earth Mover distance and motion cues. For example, the system can track a head even if the face in no more visible (e.g. hidden by hair or due to changing viewpoint). Figure 5 illustrates an accurate tracking with free-drift effect of a hand with a varying background under the guidance of optical flow as motion cues. Figure 7 illustrates the robustness of our approach in comparison to a color-based particle filter (Condensation of Perez, Hue, Vermaak, and Gangnet) that does not include our features. We show that the Condensation mixes regions having the same color shape and distribution whereas our tracker is not confused by the similar regions. This is due in particular to the addition of motion cues. Further evaluations of standard particle filter performances against existing techniques can be found in the literature (see Sect. 2).



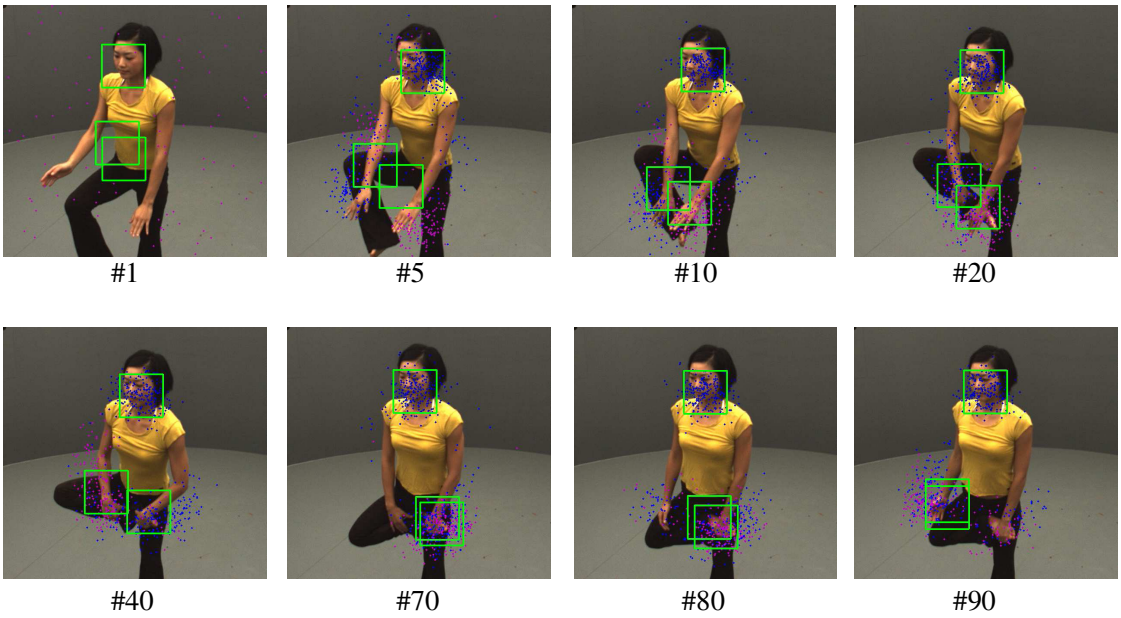
#1 #2 #3 #4

Fig. 5. Using motion cues to improve tracking. The combination of color cues and motion cues allows to perform robust tracking and prevent drift effects. The tracking of hands is efficient even with a changing background.



#1 #45 #145 #165

Fig. 6. Tracking with appearance change. The proposed approach relies on a subspace of multimodal cues (color models, motion cues, etc.) which is updated online across the video sequence. The system can track objects in motion with appearance changes.



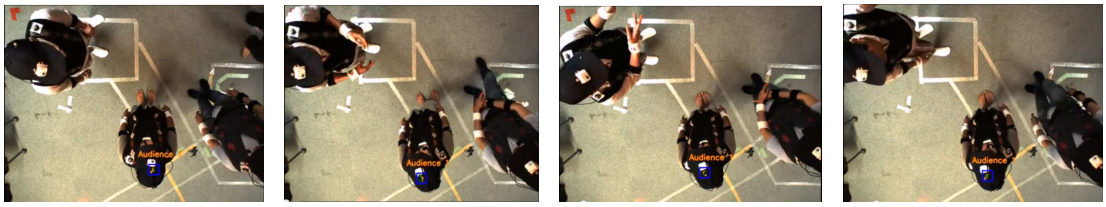
(a) Condensation.

(b) Proposed multimodal particle filter.

Fig. 7. Robust body part tracking. (a) Classical Condensation methods (Isard, & Blake, 1998; Perez, Hue, Vermaak, & Gangnet, 2002) are confused by regions with similar color and shape content. (b) In frame #20, both hands are almost included in a same tracking window, but afterwards motion cues have helped to discriminate the different tracks.

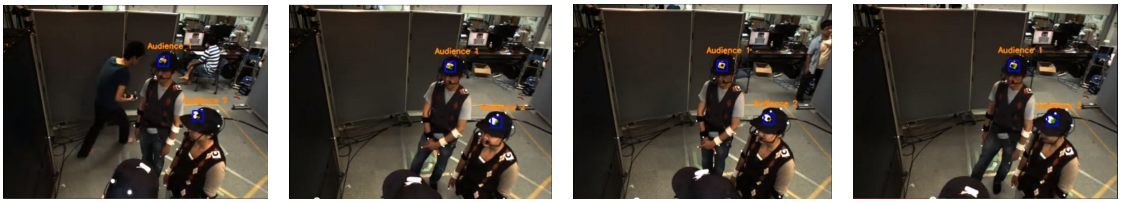
#50 #2000 #3150 #3625

(a) Camera view 1 - presenter view.



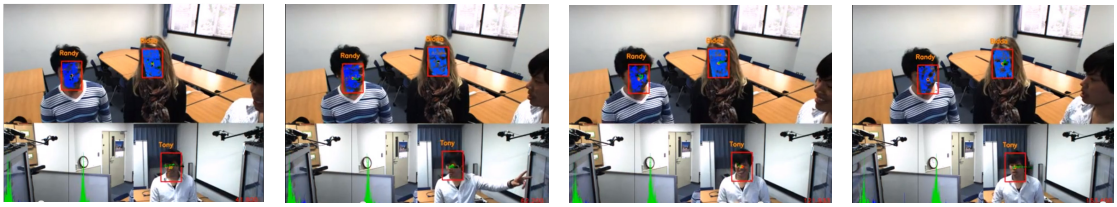
#50 #2000 #3150 #3625

(b) Camera view 2 - top view.



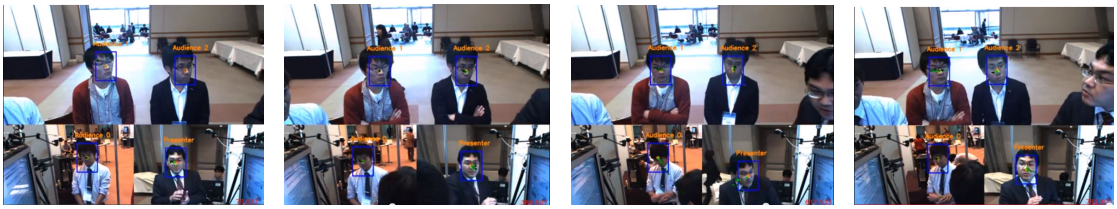
#50 #2000 #3150 #3625

(c) Camera view 3 - audience view.



#1040 #1550 #3025 #3825

(d) Multiview tracking of 3 subjects using depth cue in meeting room.



#1950 #9700 #15400 #19525

(e) Multiview tracking of 4 subjects using depth cue in conference hall (IEEE ICASSP'12 demo).

Fig. 8. Multiple people tracking from multiple views using multimodal particle filter. A presenter and its audience are tracked using a smart poster system consisting of multiple view cameras. In (a), (b) and (c), modalities include color and motion cues. In (d) and (e), modalities include color, motion, and depth. In (d), tracking system outputs are shown in red. Depth cues are represented by blue points and at bottom left (depth value distribution). In (e), tracking system outputs are shown in blue. The system tracks four subjects simultaneously in real-time.

6. CONCLUSION

Visual tracking of human body parts is a major research field due to the numerous possible applications. The literature has provided powerful algorithms based on statistical methods especially dedicated to face detection and tracking. Nevertheless, it is still challenging to handle complex object classes such as human body parts whose appearance changes occur quite frequently while in motion.
In this work, we propose to integrate multiple various cues such as color, motion and depth, in a multimodal framework based on the well-known particle filtering to leverage visual tracking efficiency. We have used the Earth Mover distance to compare color-based model distribution in the HSV color space in order to strengthen the invariance to lighting condition. Combined with an online adaptive update of multimodal template subspace, we have obtained robustness to partial occlusion. We have also proposed to integrate extracted motion features (optical flow) to handle strong appearance changes and prevent drift effect. In addition, our tracking process is run within a practical tracking-by-detection process that dynamically updates the system output. Our multimodal tracking system has been tested on real-world video data, and results on different sequences were shown. For future work, we believe our approach can be easily extended to handle a online manifold learning process. This would improve both detection and tracking processes.

ACKNOWLEDGEMENTS

This work was supported in part by the JST-CREST project "Creation of Human-Harmonized Information Technology for Convivial Society". The authors would like to thank the members of the project, as well as all the participants of the numerous capture sessions.

REFERENCES

Butt A. A., & Collins, R. T. (2013). Multi-target Tracking by Lagrangian Relaxation to Min-Cost Network Flow. *IEEE Conference on Computer Vision and Pattern Recognition.*

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

Blake, A., Curwen, R., & Zisserman, A. (1993). A framework for spatio-temporal control in the tracking of visual contours. *International Journal of Computer Vision*, 11(2), 127–145.

Bradski, G. (1998). Computer vision face tracking as a component of a perceptual user interface. *In Workshop on Applications of Computer Vision*. 214–219.

Brendel, B., Amer, M., & Todorovic, S. (2011). Multiobject Tracking as Maximum Weight Independent Set. *IEEE Conference on Computer Vision and Pattern Recognition.*

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8), 790-799

Choudhury, R., Schmid, C., & Mikolajczyk, K. (2003). Face detection and tracking in a video by propagating detection probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), 1215-1228.

Collins, R., Liu, Y., & Leordeanu, M. (2005), Online selection of discriminative tracking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 1631-1643.

Comaniciu, D., Ramesh, V., & Meeh, P. (2000). Real-time tracking of non-rigid objects using mean shift. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 142–149.

Comaniciu, D., Ramesh, V., & Meeh, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5), 564–577.

Dornaika, F., & Davoine, F. (2005). Simultaneous facial action tracking and expression recognition using a particle filter. *IEEE International Conference on Computer Vision,* 1733-1738.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3), 197–208.

Hartley, R., & Zisserman, A. (2004). Multiple View Geometry in Computer Vision Second Edition. *Cambridge University Press*.

Hillier, F.S., & Lieberman, G.J. (1990). Introduction to mathematical programming. *McGraw-Hill*.

Hjelmas, E., & Low, B.K. (2002). Face detection: a survey. *Computer Vision and Image Understanding*, 83, 236–274.

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

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Eng.*, 82, 35-45.

Kim, M., Kumar, S., Pavlovic, & V., Rowley, H. (2008). Face tracking and recognition with visual constraints in real-world videos. *IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Y., Ai, H., Yamashita, T., Lao, S., & Kawade, M. (2007). Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans. *IEEE Conference on Computer Vision and Pattern Recognition.*

Karavasilis, V., Nikou, C., & Likas, A. (2011). Visual tracking unsing the Earth Mover's distance between Gaussian mixtures and Kalman filtering, *Image Vision Computing*, 29(5), 295-305.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* , 60(2), 91–110.

Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *International Joint Conferences on Artificial Intelligence*, 674–679.

Lucena, M., Fuertes, J.M., & de la Blanca, N.P. (2004). Evaluation of three optical flow based observation models for tracking. *International Conference on Pattern Recognition*, 236–239.

Martinez, A.M., & Kak, A.C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), 228–233.

Maggio, E., Smeraldi, F., & Cavallaro, A. (2007). Adaptive multi-feature tracking in a particle filtering framework. *IEEE Transactions on Circuits and Systems for Video Technology*, 17, 10, 1348-1359.

Matsuyama, T., Nobuhara, S., Takai, T., Tung, T. (2012). 3D video and its applications. *Springer*.

Matthews, I., Ishikawa, T., & Baker, S. (2004). The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6), 810–815.

Okuma, K., Taleghani, A., de Freitas, N., Kakade, S., Little, J., & Lowe, D. (2004). A boosted particle filter: multitarget detection and tracking. *European Conference on Computer Vision*, 28–39.

Perez, P., Hue, C., Vermaak, J., & Gangnet, M. (2002). Color-based probabilistic tracking. *European Conference on Computer Vision*, 661–675.

Rehg, J., & Kanade, T. (1994). Visual tracking of high dof articulated structures: An application to human hand tracking. *European Conference on Computer Vision*, 35–46.

Ross, D., Lim, J., Lin, R., & Yang, M. (2008). Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1), 125-141.

Rubner, Y., Tomasi, C., & Guibas, L.J. (1998). A metric for distributions with applications to image databases. *IEEE International Conference on Computer Vision*, 59–66.

Shirdhonkar, S., & Jacobs, D.W. (2008). Approximate Earth mover's distance in linear time. *IEEE Conference on Computer Vision and Pattern Recognition.*

Sugimoto, A., Yachi, K., & Matsuyama, T. (2003). Tracking human heads based on interaction between hypotheses with certainty. *The 13th Scandinavian Conference on Image Analysis*.

Terzopoulos, D., & Szeliski, R. (1992). Tracking with Kalman snakes. In *Active Vision*, A. Blake and A. Yuille (Eds.), MIT Press: Cambridge, MA, pp. 3–20.

Tola, E., Lepetit, V., & Fua, P. (2008). A fast local descriptor for dense matching. *IEEE Conference on Computer Vision and Pattern Recognition*.

Tomasi, C., & Kanade, T. (1991). Detection and tracking of point features. *Technical Report CMU-CS-91-132*, Carnegie Mellon University.

Tung, T., & Matsuyama, T. (2008). Human Motion Tracking using a Color-Based Particle Filter Driven by Optical Flow. *European Conference on Computer Vision Workshop.*

Tung, T., & Matsuyama, T. (2012). Topology Dictionary for 3D Video Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8), 1645-1657.

Tung, T., Gomez, R., Kawahara, T., & Matsuyama, T. (2012). Group Dynamics and Multimodal Interaction Modeling using a Smart Digital Signage, *European Conference on Computer Vision Workshop.*

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition*, 511–518.

Wang, J., Chen, X., & Gao, W. (2005). Online selecting discriminative tracking features using particle filter. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 1037-1042.

Wang. X., & Tang, Z. (2010). Modified particle filter-based infrared pedestrian tracking, *Infrared Physics & Technology*, 53, 4, 280-287.

Wu, Y., Lim, J., & Yang, M. H. (2013). Online Object Tracking: A Benchmark, *IEEE Conference on Computer Vision and Pattern Recognition*.

Yilmaz, A., Javed, O., & Shah, M., (2006). Object tracking: A survey. *ACM Computer Survey*, 38, 4, 1-45.