2-2007

Moving-object detection, association, and selection in home videos

Zailiang PAN

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Citation

1

Moving Object Detection, Association and Selection in Home Videos

Zailiang Pan, Chong-Wah Ngo *Member, IEEE*

***Abstract*— Due to the prevalence of digital video camcorders, home videos have become an important part of life-logs of personal experiences. To enable efficient video parsing, a critical step is to automatically extract objects, events and scene characteristics present in videos. This paper addresses the problem of extracting objects from home videos. Automatic detection of objects is a classical yet difficult vision problem, particularly for videos with complex scenes and unrestricted domains. Compared with edited and surveillant videos, home videos captured in uncontrolled environment are usually coupled with several notable features such as shaking artifacts, irregular motions and arbitrary settings. These characteristics have actually prohibited the effective parsing of semantic video content using conventional vision analysis. In this paper, we propose a new approach to automatically locate multiple objects in home videos, by taking into account of *how* and *when* to initialize objects. Previous approaches mostly consider the problem of *how* but not *when* due to the efficiency or real-time requirements. In home video indexing, online processing is optional. By considering *when*, some difficult problems can be alleviated, and most importantly, enlightens the possibility of parsing semantic video objects. In our proposed approach, the *how* part is formulated as an object detection and association problem, while the *when* part is a saliency measurement to determine the best few locations to start multiple object initialization.**

I. Introduction

Nowadays, home video is commonly used to document personal and family events. The management of home videos, nevertheless, is an extremely tedious task even with the aid of manual editing tools such as Adobe Premiere [1] and Apple iMovie [2]. Very often, the close examination of video data on a frame-by-frame basis is required to parse the useful content in videos. The "search of personal memories" from a set of unedited, long-winded and low quality videos becomes harder without proper content parsing and indexing. Recently, automatic content analysis of home videos has indeed attracted numerous research attention due to its commercial potential in providing automatic browsing, summarization and editing functionalities. In the past decade, numerous research has been conducted for content-based representation and analysis, but mainly for the professionally edited videos such as movies, news and sport videos. Relatively little work has been dedicated to the domain of home videos. Home videos, unlike scripted and edited videos, are domain unrestricted and usually interwoven with undesirable shaking artifacts [3]. These characteristics indeed prohibit the effective parsing of home videos and present the new aspect of technical challenges for content analysis.

Existing research areas in home video analysis include scene analysis [4], [3], browsing [5], [6], content structuring [7], highlight detection [8], summarization [9], [10] and editing [11]. In these papers, camera motion [8], face [6], time stamp [9], [10] and temporally order information [3] are frequently exploited for the content organization of home videos. While most of the work addresses the issues of video abstraction [4], [3], [8], [9], [10] and browsing [5], [6], relatively little work has been conducted on the detection of video objects for parsing and indexing [12], [13], [14]. Video objects play a special role in personal life-logs since they can be any symbolically memorable object such as people, buildings or notable patterns that most queries may engross. Besides keyframe, snippet [7], shot, scene [4], [3] and event [9], [10] representation, perhaps another way of indexing home videos is to "album" the content with object patterns to facilitate searching and browsing.

A fundamental problem in extracting and indexing video objects is to initialize the meaningful and huge variety of objects that are concealed in a bunch of jerky frames with unknown visual quality. This paper addresses the issues of offline multiple object initialization by considering the situations of *how* and *when* to initialize. Traditionally most approaches in object extraction assume good initial conditions (e.g., number, positions and appearances of objects) can always be obtained through information training or from the first few frames. This assumption, nevertheless, is invalid since home videos can be captured anywhere and anytime without a specific setting or environment. Moreover, due to the amateur operation of camcorders, the portions of videos appropriate for object initialization are random in nature. Intuitively, we should initialize objects at locations where semantic objects can be integrally segmented and then start information association to disambiguate data observation. Inspired by this idea, we propose a three-step approach grounded on detection, association and selection to tackle this problem. In detection phase, a bag of candidates, which represents the current observation on hand, is detected. In selection phase, the best possible candidates, which round up the good initial conditions of objects, are drawn from the bag with certain degree of confidence. In association phase, up-to-date evidence and recent observation are adjoined to update object evolution. The major contribution of our work is that both "when and how to

Manuscript received July 1, 2005; revised Nov 26, 2005. The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 1072/02E).

Z. Pan and C. -W. Ngo are with the City University of Hong Kong.

C. -W. Ngo is the contact author. Please direct all correspondences by email: cwngo@cs.cityu.edu.hk.

initialize" are jointly considered to allow collaborative and robust initialization of objects in home videos.

The remaining paper is organized as follows. Section II describes the existing techniques in object detection, tracking and initialization. Section III presents the overview of our proposed approach. Section IV describes our approach in detecting object candidates. Based on 3D tensor representation, robust motion estimation and clustering algorithms are proposed to detect the bag of candidates in video shots. Section V presents the association of objects under the framework of recursive Bayesian filter. Kalman filter is employed for the data association of initial object condition. Section VI describes our algorithm in selecting the best possible candidates for bi-directional association along the temporal dimensional. Finally, Section VII presents experimental results and Section VIII concludes this paper.

II. Related Works

Object detection and tracking have been extensively studied in the past few decades [15], [16], [17], [18]. Both tasks need the initialization of parameters such as the sizes, positions, appearance, and numbers of objects. In [19], Weiss and Adelson incorporate spatial coherence constraint in a motion mixture framework to estimate the number of objects, but introduce an extra parameter to prescribe the accuracy of a mixture model. In [16], Sawhney and Ayer propose minimum description length (MDL) to determine the number of models. A similar model selection method based on minimum message length (MML) like criterion is also presented in [20]. Both MDL and MML are based on information theory, and rely on the trade-off of the message code length between model parameters and model data. However, the message length of model data is sensitive to noise since the length can change dramatically due to outlier which is a common phenomena in home videos. Indeed, if the frames with the best possible initial conditions of objects can be effectively discovered, the overhead of dealing with noises and model selection strategies can be significantly alleviated.

Object initialization is also required during and after occlusion. Popular approaches include the utilization of multiple cameras for occlusion analysis, and multiple hypothesis tracking (MHT) which generates, maintains and prunes hypotheses over time based on object dynamics. For instance, Mittal and Davis use multiple synchronized cameras to track multiple people to avoid the lack of visibility when occlusion happens [21]. Yeasin *et al.* employs MHT to investigate the coherency of object trajectory [22]. The complexity of MHT, nevertheless, grows dramatically as the number of hypotheses increases, which causes MHT an computationally expensive approach. In this paper, we do not consider the case of multi-camera since home videos are often captured by single cameras. In addition, MHT is not adopted due to speed consideration, and the fact that MHT requires additional mechanism for hypothesis pruning which cannot be easily dealt with when considering the shaking artifacts in home videos.

The hyper-linking, mining and extraction of video objects have been addressed in [12], [13], [14], [7]. In [12], an approach to hyper-link objects of interests is proposed. The objects are manually specified and then used as exemplars for candidate localization. A metric mixture model framework is formulated to model the joint probabilistic of exemplars and their geometric transformations in a novel configuration space. Candidate objects are then localized in the space with importance sampling for hyper-linking. The works in [13], [14] investigate the automatic matching, mining and extraction of objects with text retrieval approach. In [13], a vocabulary composed of visual words is built for content representation. Each word corresponds to a vector-quantized viewpoint invariant region descriptor. With this vocabulary, keyframes are described by a set of visual words and represented with a vector space model. Given an user-defined object query, keyframes containing similar objects can be rapidly searched with the use of inverted files as in text based retrieval. In [14], the approach in [13] is further extended for object mining by measuring the re-occurrence of spatial configurations of visual words in keyframes. A critical step in this approach is the definition of spatial configuration to enable the mining capability. Both [14] and our approach are unsupervised and with the ultimate aim of extracting key objects from videos. In term of proposed techniques, a fundamental difference between these two works is that [14] is appearance driven, with visual words as descriptors and spatial configurations as constraints to guide object extraction. Our approach, on the other hand, is motion driven, and the goal is to mine the best possible object candidates from frames to drive object initialization. Similar in spirit, we propose a novel seed selection (NSS) algorithm in [7] for finding good candidates to start initialization. The works presented in this paper are different from [7] in the following aspects: i) the selection of multiple seeds are supported, ii) the selection and association of seeds are jointly considered, iii) concrete experiments are conducted to verify the effectiveness of seed selection. In [7], only single seed selection is considered, and the problem of data association is not addressed.

III. Overview of Proposed Approach

While most approaches consider *how* to initialize, we investigate *how* and *when* to initialize the variety of objects in videos. Figure 1 depicts the major components in our approach. The *when* part is handled by temporal selection, while the *how* part is based on data association. We term the initial condition of an object in a frame as a seed, and the measurements as seed candidates. A seed is modeled as a state, and it forms a state sequence to describe the evolution of an object along the temporal dimension. Initially, a bag of seed candidates is collected at each video shot, as shown in the bounding boxes of Figure 1(b). The initial parameters (sizes, positions, appearance and object numbers) are estimated by robust motion and cluster analysis. These candidates represent the varying conditions of objects at different frames, where some appear in their entirety while others appear in fragmented forms. A saliency plot, as shown in Figure 1(c), is then computed to model the fidelity of candidates. Relying on this measure, temporal seed selection draws the best

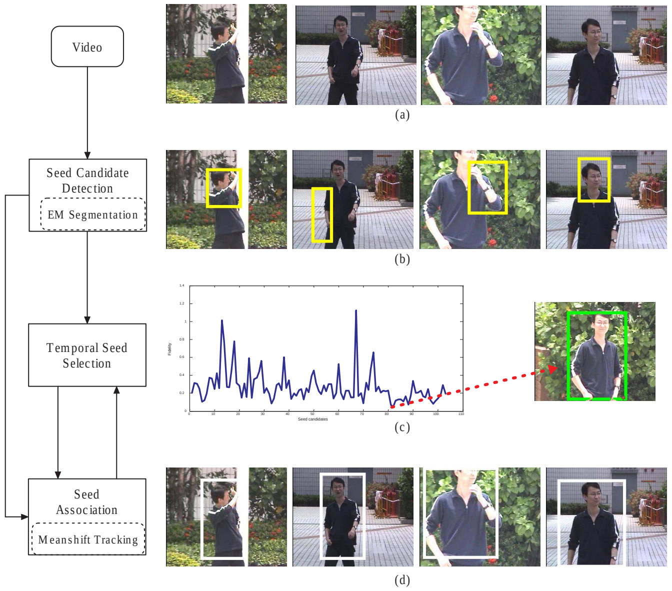


Fig. 1. Proposed framework. (a) Four sample video frames in a shot; (b) The bag of detected candidates indicated by the bounding boxes in the sample frames; (c) The first seed (right) is selected based on the computed saliency plot (left) with y-axis shows the fidelity of seed candidates in the shot; (d) The result of extracting objects after propagating and associating the first seed in (c) with the candidates at (b).

possible candidates from the bag of candidates to activate seed association. In Figure 1(c), a seed is selected and then associated with the candidates in (b) to produce the results in (d). Basically, the seed selection and association are geared sequentially over time to continuously associate and select seeds as appropriate to guarantee the integrality of objects found in the bag, as shown in Figure 1(d). Seed association is grounded on recursive Bayesian filter and deals with the temporal dynamics (including changes of size, position and appearance) of objects.

In Figure 1, notice that EM segmentation and meanshift tracking algorithms are two optional components in seed candidate detection and association. The role of EM is to refine the segmented regions of detected candidates, while the role of meanshift is to predict the temporal changes of seeds. Indeed, most existing segmentation and tracking algorithms can be directly employed in this framework. We choose EM and meanshift algorithms due to the consideration of robustness and efficiency.

A. *Notation*

In the remaining sections, a seed is denoted by s_t^j, where the subscript t denotes the time and the superscript j denotes the seed of a j^{th} object at time t. For simplicity, we omit the symbol j or t, unless necessary. In other words, we let s_j means the seed of j^{th} object at time t and let s_t indicate a seed at frame t. Furthermore, we denote $\{s_t\}$ as the set of seeds in a shot and $\{s_t^j\}_j$ as a set of seeds at time t. The notations c_t^j, c_t, c_j, $\{c_t\}$ and $\{c_t^j\}_j$ for seed candidates are defined in the same way.

B. *Seed Parameterization*

The aim of parameterization is to define a concrete representation for seeds. There are three major concerns in the parameterization:

1) Dependency. In this paper, we use EM and meanshift algorithms. The common initial condition for both algorithms is the support layer of an object pattern. Thus it is straightforward to base the seed definition on the object region.
2) Simplicity. In general, complex seed parameterization can lead to accurate representation but with the expense of a heavy computational load. Seeds represent the initial conditions and thus do not require precise representation, therefore a simple form of parameterization is preferred for ease of processing.
3) Linearity constraint. Since we use Kalman filter (Section V-B), the seed parameters should be represented by a vector of real values. Note that this constraint is not compulsory if other state space approach is used.

Based on the concerns, we use a rectangle region to represent a seed. A seed state is defined by the following vector,

$$s_t = [m_x, m_y, r_x, r_y]^T \tag{1}$$

where $m = [m_x, m_y]^T$ is the upper-left corner of the seed rectangle, while $r = [r_x, r_y]^T$ is the opposite corner, such that $m_x \leq r_x$ and $m_y \leq r_y$. To simplify the measurement model of seed candidate, the state vector of a candidate is defined similarly.

With reference to Figure 1, seed candidate detection based on motion estimation and clustering will be described in the next section. The association of seeds under the framework of recursive Bayesian filter will be outlined in Section V, while the sequential selection of seeds will be presented in Section VI.

IV. Seed Candidate Detection

We exploit motion cues for the detection of seed candidates. Tensor representation is employed to compute optical flows and their associated fidelity values, while robust motion clustering is proposed to effectively discover the number of seed candidates.

A. *3D Tensor Representation*

Let $I(x, y, t)$ be the space-time intensity of a point in a $3D$ image volume. Assume $I(x, y, t)$ remains constant along a motion trajectory, optical flow is computed as

$$\frac{dI}{dt} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = \epsilon \tag{2}$$

where u and v represent the components of local spatial velocity, and ϵ is a noise variable assumed to be independent, white and zero-mean Gaussian. Eqn (2) is the inner product of a homogeneous velocity vector V and a spatio-temporal gradient ∇I, i.e.,

$$(\nabla I)^T V = \epsilon \tag{3}$$

where $V = [u, v, 1]^T$ and $\nabla I = \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t}\right]^T$. The noise term ϵ^2 can be used as a fidelity measure for motion estimation [23]. Nevertheless, any fidelity measure that depends only on ϵ cannot fully exploit the fact that the estimated local velocity in a region with high intensity variability is more reliable than in a region with low variability. To tackle this problem, we introduce a fidelity term based on $3D$ tensor representation for robust estimation. Under the assumption that the flows are constant over a $3D$ volume R, the total sum of ϵ^2 in R can be derived as

$$E = \sum \epsilon^2 = V^T \left(\sum_{x,y,t \in R} (\nabla I)(\nabla I)^T \right) V \qquad (4)$$

The central term is a symmetric tensor which represents the local structure of R in space-time dimension. The tensor has the form

$$\Gamma = \begin{bmatrix} J_{xx} & J_{xy} & J_{xt} \\ J_{yx} & J_{yy} & J_{yt} \\ J_{tx} & J_{ty} & J_{tt} \end{bmatrix} \qquad (5)$$

where

$$J_{mn} = \sum_{x,\,y,\,t \in R} \frac{\partial I}{\partial m} \frac{\partial I}{\partial n} \quad m, n = x, y, t$$

Given the tensor representation in Eqn (5), the optical flows can be estimated by minimizing the cost function E in Eqn (4). The diagonal components of a tensor which represent the intensity variation in spatio-temporal coordinate can be exploited for fidelity measure. Thus, our proposed fidelity term λ, which depicts the certainty of estimated optical flow in R, is defined as

$$\lambda = 1 - \frac{E}{E + J_{xx} + J_{yy}} \qquad (6)$$

The fidelity term has following favorable properties: i) it is maximal for ideal flows, i.e., $E = 0$, ii) it is minimal if no spatial intensity variation, i.e., $J_{xx} + J_{yy} = 0$, iii) its value is normalized in the range $[0, 1]$, since $J_{xx} + J_{yy} \geq 0$ and $E \geq 0$.

B. Robust Clustering

Given the optical flows $\{\mathbf{v}_i\}$ and their fidelities $\{\lambda_i\}$ (Eqn (6)) at time t, we employ k-means algorithm to cluster optical flows in each frame. The number of clusters, g, is initially set to a reasonably large value[1]. The clusters are subsequently merged one by one based on the pairwise distance between clusters. The detailed algorithm is given in Algorithm 1. The robustness arises from two aspects: i) weight the importance of $\mathbf{v}_i$ with λ_i, ii) use of MVE robust estimator [24]. By exploiting the clustering indices $\{u_{ij}\}$ (Algorithm 1) and the corresponding spatial positions $\{\mathbf{p}_i\}$ of the optical flows $\{\mathbf{v}_i\}$ in frame t, a seed candidate c_j is represented by its centroid $\bar{\mathbf{p}}_j$ enclosed under a rectangle with size $\mathbf{d}_j$ as follows

$$c_j = \begin{bmatrix} (\bar{\mathbf{p}}_j - \mathbf{d}_j) \\ (\bar{\mathbf{p}}_j + \mathbf{d}_j) \end{bmatrix} \quad \forall j \in 1 \cdots g \qquad (7)$$

Algorithm 1 Robust motion clustering

```
1) Given a cluster number $g$, $k$-means is used to compute the initial cluster matrix $\{u_{ij}\}$, where $u_{ij} = 1$ if $\mathbf{v}_i \in j^{th}$ cluster and $u_{ij} = 0$ otherwise.
2) Calculate the cluster probability $p_j$ by
       $p_j = \frac{\sum_i u_{ij}}{\sum_j \sum_i u_{ij}}$
3) Compute the cluster mean $\mathcal{M}_j$ and covariance matrix $\mathcal{C}_j$ by the robust estimator, Minimum Volume Ellipsoid (MVE) in [24].
4) Compute the distance $d_{kl}$ between clusters $k$ and $l$
       $d_{kl} = s(1-s)(\mathcal{M}_k - \mathcal{M}_l)^T [s\mathcal{C}_k + (1-s)\mathcal{C}_l]^{-1} (\mathcal{M}_k - \mathcal{M}_l)$
   where $k, l \in 1 \cdots g$ and $s = p_k/(p_k + p_l)$.
5) Select two clusters $k^*$ and $l^*$, $k^* < l^*$, such that $d_{k^*l^*} = \min_{ij}\{d_{ij}\}$
6) If $d_{k^*l^*} < 2.0$, merge both clusters and set $g = g - 1$.
7) Terminate if no merge of clusters, else go to step 2.
```

where

$$\bar{\mathbf{p}}_j = \frac{\sum_i (\mathbf{p}_i u_{ij})}{\sum_i u_{ij}}$$

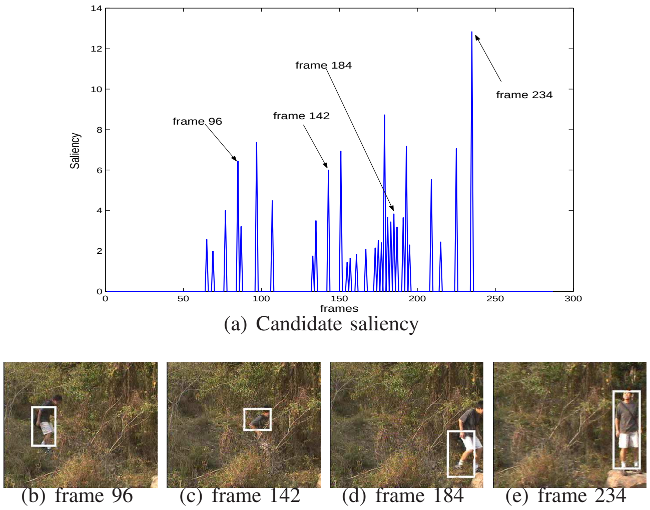
$$\mathbf{d}_j = \sqrt{\frac{\alpha \times diag\{\sum_i (u_{ij}(\mathbf{p}_i - \bar{\mathbf{p}}_j)(\mathbf{p}_i - \bar{\mathbf{p}}_j)^T)\}}{\sum_i u_{ij}}}$$

$u_{ij} = \{0, 1\}$ specifies the membership of a optical flow, and $\mathbf{d}_j$ is computed based on the standard deviation of cluster j. The constant $\alpha = 3$ and this value is calculated by comparing the ratio of variance to the half-length square of a univariate uniform distribution. The saliency of candidates η at frame t is measured based on cluster separability:

$$\begin{aligned} \eta &= tr(\eta_w^{-1} \eta_b) \qquad (8) \\ \eta_w &= \sum_{j=1}^{g} p_j \mathcal{C}_j \\ \eta_b &= \sum_{j=1}^{g} p_j (\mathcal{M}_j - \sum_{k=1}^{g} p_k \mathcal{M}_k)(\mathcal{M}_j - \sum_{k=1}^{g} p_k \mathcal{M}_k)^T \end{aligned}$$

where η_w and η_b are the intra and inter cluster distances respectively [25]. In Eqn (8), p_j, $\mathcal{M}_j$, $\mathcal{C}_j$ are respectively the probability, mean and covariance of cluster j (details in Algorithm 1). Basically, a large value of η_b and a small value of η_w is favored since overall they hint the confidence of extracting objects in their entirety. To demonstrate the use of saliency measurement, Figure 2(a) shows the values of η in a sequence. In this sequence, the appearance of target object undergoes drastic changes due to occlusion, camera motion and $3D$ object motion. Figures 2(b)-(e) show the detected candidates at four frames with different η values. Obviously, it is more appropriate to start initialization at frame 234 which holds the highest η in Figure 2(e) rather than at frames 96, 142 and 184.

[1] Typically there are only few moving objects (about 1 to 4) in a frame. In the experiment, g is deliberately set to 8 to tackle the extreme case when there are twice than the number of expected objects.



(a) Candidate saliency

(b) frame 96 (c) frame 142 (d) frame 184 (e) frame 234

Fig. 2. Candidates with different saliency

C. EM algorithm

The candidates obtained through robust clustering are indeed rough approximations of object location. To refine the boundary of seed candidates, we modify EM algorithm in [16] for finer candidate segmentation. Different from [16], our EM algorithm takes the initial conditions (e.g., number of candidates, approximate regions of seed candidates) directly from Section IV-B as inputs. We modify the E-step of [16] so that the conditional expectation takes into account the prior initial conditions of objects at each iteration. As a result, our EM algorithm is more efficient and stable than [16] owing to the fact that we do not treat each pixel as an independent element during optimization. The original EM algorithm is likely to merge regions with relatively small size than other regions, while the modified version can deal with this problem effectively.

V. Seed Association

We model a seed sequence with a discrete-time dynamic system and use the state-space approach for seed estimation. Denote $S_t = \{s_0, \cdots, s_t\}$ as the seed history, the goal is to estimate a seed s_t, given all its candidates $C_t = \{c_0, \cdots, c_t\}$ up to that moment. This is equivalent to construct a posterior probability density function $p(s_t|C_t)$. Assuming the dynamic system forms a temporal Markov chain as follows

$$p(s_t|S_{t-1}) = p(s_t|s_{t-1}) \tag{9}$$

The seed candidates are further assumed to be independent, both mutually and with respect to the seeds, or equivalently

$$p(C_t|S_t) = \prod_{i=0}^{t} p(c_i|s_i)$$

The theoretical optimal seed estimation can then be described by a recursive Bayesian filter [17] as

$$p(s_t|C_t) = k_t p(c_t|s_t) p(s_t|C_{t-1}) \tag{10}$$

where

$$p(s_t|C_{t-1}) = \int_{s_{t-1}} p(s_t|s_{t-1}) p(s_{t-1}|C_{t-1})$$

and k_t is a normalization constant that does not depend on s_t. There are two steps involved in this propagation. The *prediction* step employs the seed posterior at time $t-1$, $p(s_{t-1}|C_{t-1})$, to derive the seed prior at time t, $p(s_t|C_{t-1})$, through the seed dynamic $p(s_t|s_{t-1})$. Then the *update* step uses the seed candidate density $p(c_t|s_t)$ to compute the seed posterior at time t in Eqn (10).

A. Seed and Candidate Modeling

The aim of seed estimation is to guarantee accurate object initialization by associating all the seeds that belong to an object. To predict prior seed density, we model the seed dynamic as

$$s_t = f(s_{t-1}) + v_t \tag{11}$$

where $f(s_{t-1})$ can be any tracking algorithm and v_t is the prediction noise which is assumed to be independent, white and with zero-mean Gaussian density, i.e., $p(v_t) \sim N(0, V_t)$. The choice of noise covariance V_t is related to the tracking algorithm being used. Combining Eqn (9) and Eqn (11), we have

$$p(s_t|s_{t-1}) \propto exp(-\frac{1}{2}(s_t - f(s_{t-1}))^T V_t^{-1}(s_t - f(s_{t-1}))) \tag{12}$$

A candidate model describes the relationship between a seed s_t and its observation c_t. We construct the model $p(c_t|s_t)$ as

$$c_t = s_t + u_t \tag{13}$$

where u_t is the measurement noise with $u_t \sim N(0, U_t)$ and U_t is the covariance of noise used to describe the uncertainty of a seed candidate. Since the saliency η_t (Eqn (8)) represents the confidence of seed candidate estimation, it is used to form the uncertainty U_t,

$$U_t = \frac{1}{\eta_t} I$$

where I is a 4×4 identity matrix. To be consistent with Eqn (10), Eqn (13) is re-written as

$$p(c_t|s_t) \propto exp(-\frac{1}{2}(c_t - s_t)^T U_t^{-1}(c_t - s_t)) \tag{14}$$

B. Seed Filtering

Combining Eqn (10), Eqn (12) and Eqn (14), Kalman filter is used for the data association of initial conditions. The prediction and update steps are given as follows

- Seed predict
$$\begin{aligned} \tilde{s}_t &= f(\hat{s}_{t-1}) \\ \tilde{P}_t &= g(P_{t-1}) + V_t \end{aligned} \tag{15}$$
- Seed update
$$\begin{aligned} K_t &= \tilde{P}_t(\tilde{P}_t + U_t)^{-1} \\ \hat{s}_t &= \tilde{s}_t + K_t(c_t - \tilde{s}_t) \\ P_t &= (I - K_t)\tilde{P}_t \end{aligned} \tag{16}$$

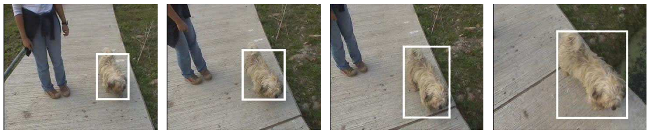


Fig. 3. Tracking with changes of scale and appearance.

where $\tilde{s}_t$ is the prior seed state at time t, $\hat{s}_t$ is the posterior estimated seed state, $\tilde{P}_t$ is prior error covariance, P_t is posterior error covariance, and K_t is the Kalman gain. In the prediction step, we use $g(P_{t-1})$ to represent the error propagation of seed tracking. Through seed predict and update, we establish a framework for seed association. The prediction mainly relies on tracking algorithm, while the update can adapt a seed state to the changes of size, position and appearance.

C. Seed Tracking

We adopt meanshift algorithm in [18] for seed tracking. The similarity measure of meanshift is based on Bhattacharyya coefficient metric between the color density distributions of a target model and a target candidate (in the form of color distribution). By incorporating meanshift tracker into seed association, the $f()$, $g()$ and V_t in Eqn (15) can be defined. For tracking, $\tilde{s}_t = f(\hat{s}_{t-1}) = \hat{s}_{t-1} + d_t$, where d_t is the seed displacement estimated by meanshift tracking algorithm. Based on the seed parameterization in Eqn (1), both $\tilde{s}_t$ and $\hat{s}_t$ are represented in the form of $[(\mathbf{y} - h)^T, (\mathbf{y} + h)^T]^T$ in meanshift tracking, where $\mathbf{y}$ is the position of target candidate and h is the radius of kernel profile [18]. In addition, we assume that the prediction process by Eqn (15) does not produce process noise. Therefore $\tilde{P}_t = g(P_{t-1}) + V_t = P_{t-1}$, and consequently the Gaussian prediction in Eqn (12) is simplified to deterministic model which can still be represented with Kalman filter (see Appendix I for proof). Figure 3 shows the results of tracking by incorporating meanshift in Kalman filter. Instead of enumerating the possible sizes of an object as in [18], our tracker can automatically adapt to the changes of scale and appearance through the prediction and updating steps.

VI. Temporal Selection of Seeds

Although seed association presents an effective way of estimating seeds from seed observations, it is still not enough for robust object initialization due to the following reasons.

1) *Intermittent motion* (IM). Certain objects may be in the state of cease-move-cease. It is difficult to automatically initialize those objects at certain frames. Multiple-frame analysis can probably solve this problem. However, it is a difficult task for the situation where camera motion and scene structure are complex and not under control.
2) *Occlusion* (OO). When objects are occluded, their visual features are not observed and thus tracking cannot continue. One popular way is to predict object movement during occlusion. However, the updating of object appearance after occlusion remain a difficult problem.
3) *New object* (NO). A mechanism is required to alert the appearance of new objects. This circumstance is somewhat similar to case 1 (intermittent motion). Both cases are actually the problem of when to start the initialization.
4) *Jerky and unstable motion* (JM). Cases 1-3 are even more difficult when jerky and unstable camera motions occur. Obviously, it is not appropriate to initialize objects at frames with shaking artifacts.

In other words, object initialization involves not only *how* but also *when* to initialize. In this section, we present our techniques on *when* to select the best seeds for bi-directional association. In addition, we discuss how to integrate seed selection and association in an elegant framework for the aforementioned problems.

A. Single Seed Selection

To determine whether a seed is better than others along the time axis, we define seed saliency based on the prior seed candidate probability density $p(c_t|s_t)$ in Eqn (14). Through Bayesian rule, $p(s_t|c_t) \propto p(c_t|s_t)p(s_t)$. It is reasonable to assume $p(s_t)$ to be uniform distribution since a seed may appear anywhere in a frame, then

$$p(s_t|c_t) \propto p(c_t|s_t). \tag{17}$$

So the conditional expectation of a seed s_t, $\bar{s}_t$, is given by

$$\bar{s}_t = E(s_t|c_t) = \int_{s_t} s_t p(s_t|c_t) \tag{18}$$

In Eqn (14), $p(c_t|s_t)$ follows Gaussian distribution. Since $p(s_t|c_t) \propto p(c_t|s_t)$, Eqn (18) is implemented with $\bar{s}_t$ being assigned to a candidate c_t. The probability of selecting a good quality seed at time t, based on the covariance of $\bar{s}_t$, is defined as

$$F_t = F(\bar{s}_t) = E((s_t - \bar{s}_t)^T(s_t - \bar{s}_t)). \tag{19}$$

We name F_t the fidelity of a seed. The value of F_t indicates the confidence of conditional seed estimation. Combining Eqs (14), (17)-(19), F is inversely proportional to η in Eqn (8). Normally, the quality of initial conditions can be assured if the best seed is selected as the one to start seed association. Thus the first seed is selected at

$$s^* = \arg\min_{\bar{s}_t}(F(\bar{s}_t)). \tag{20}$$

Once a seed s^* is selected, the seed association of s_t is activated along the time axis. Given a set of seed candidates $\{c_t^j\}_j$ at time t, the candidate used to update the seed s_t is selected as

$$c_t^* = \arg\max_j \rho(c_t^j, s_t) \tag{21}$$

where

$$\rho(c_t^j, s_t) = \frac{area(s_t \cap c_t^j)}{area(c_t^j)} \tag{22}$$

specifies the normalized overlapped regions of s_t and c_t^j. We select a candidate with the largest overlapping region for seed updating in Eqn (16).

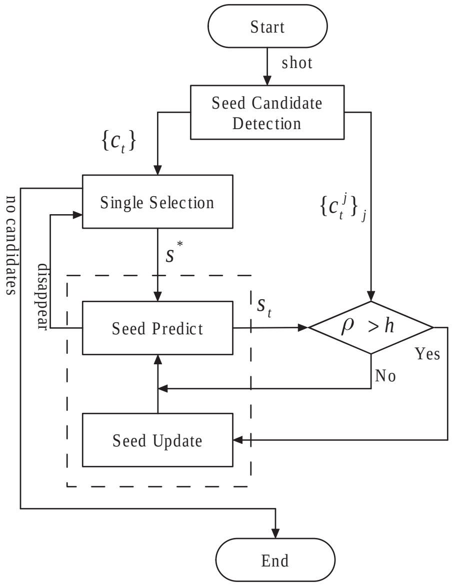


Fig. 4. Seed selection + association

B. Sequential Selection and Association

Temporal seed selection is indeed a multiple scan of single selection in Eqn (20), which is activated one after another to fully utilize all the seed candidates. The whole process is depicted in Figure 4. After candidate detection, the seed selection and association are activated sequentially. In the prediction step, if an object disappears (out of scene), another seed will be drawn from bag. This procedure repeats until all candidates are visited. Notice that an empirical threshold h is set to determine the ownership of each c_t^j after seed prediction. If $\rho > h$ (in Eqn (22)), the candidate c_t^j is used to update its corresponding seed s_t based on Eqn (21). Otherwise, no seed update is performed. The empirical threshold h is used to "gate" the updating of seeds. For robust tracking, the value of h should not be too low to avoid arbitrary updating. On the other hand, a higher value of h can cause over bagging of seed selection. In our experiment, we simply set $h = 0.5$ which compromises both extreme cases.

Algorithm 2 describes the algorithm in detail. A new vector variant b_t, seed visiting indicator, is introduced to record the history of seed visiting. In seed association, if a seed candidate c_t belongs to an object and is used for seed update, c_t is said visited and we set the corresponding visiting indicator $b_t = 1$. Initially, all the seed candidates are detected in a shot. Then the single selection is used to select the first and best seed to start a procedure of seed association. This seed association proceeds until an object cannot be tracked[2]. The single selection is invoked again to start another seed association based on the remaining unvisited seeds.

Algorithm 2 Temporal seed selection

```
1) Detect the seed candidates $\{c_t\}$ in the shot and estimate the conditional pdf $\{p(c_t|s_t)\}$.
2) Calculate the expected seeds $\{\bar{s}_t\}$ by Eqn (18) and the associated saliency measurements $\{F_t\}$ through Eqn (19). Set seed visiting indicator $b_t = 0, \forall t$.
3) Select the seed with the maximum saliency $\{F_t\}$ to be the first seed for association, $\forall t$ where $b_t = 0$.
4) Invoke seed association by Eqn (15) and Eqn (16). The seed candidates used for seed update are selected by Eqn (21)-(22). If a seed candidate $c_t$ is selected, set $b_t = 1$.
5) Go to Step 3 until no seed candidates.
```

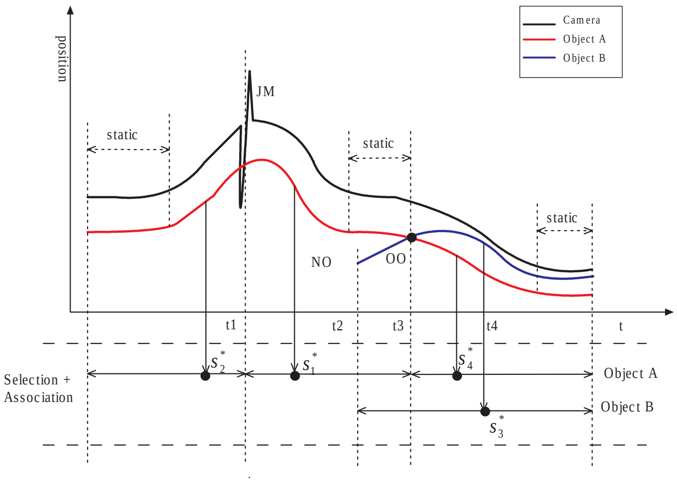


Fig. 5. Offline object initialization (OO: occlusion, NO: new object, JM: jerky camera motion)

Figure 5 shows an example to illustrate how our approach handles the difficulties when two objects which appear at different time stamps of a jerky sequence occlude each other. The challenges encountered in this sequence include jerky camera motion, occlusion, appearance of new object, and intermittent motion. To better explain the idea, we only show one spatial dimension in this example. The horizontal and vertical axes represent the time and spatial dimension respectively. At the beginning of this sequence, camera (the thick black curve) tracks an object A (the thick red curve) and causes shaking artifact at time t_1. The object A undergoes a series of intermittent motion and rests at three different time frames. A new object B which appears at time t_2 occludes object A at time t_3. Based on our algorithm, suppose the first seed selected by Eqn (20) is s_1^*. With seed association, this seed is temporally propagated in the forward and backward directions, until at time t_1 and t_3 due to camera shaking and occlusion respectively. As a consequence, two new seeds, s_2^* and s_4^*, are selected for further association. In addition, a new seed s_3^* which represents the best initial condition of object B in this sequence is found at time t_4 for association. In brief, by sequentially selecting the seeds s_1^*, s_2^*, s_3^* and s_4^*, the two objects are properly initialized throughout the shot.

VII. Experiments

We conducted experiments on eleven home videos. The first two videos (*lgerca_lisa_1* and *lgerca_lisa_2*) are from MPEG-7

[2] We use the Bhattacharyya coefficient of meanshift which indicates the appearance degradation of an object to determine whether to stop tracking.

standard test set. The others are home videos collected from different people. These videos include clips on outdoor expedition, student activity, harbor and park. In Table I, we name each video according to physical location, object or activity. These videos are composed of varying indoor and outdoor scenes, and with undesirable features of shaking artifacts, motion blurs and illumination changes. Each video is initially partitioned into shots. For *lgerca_lisa_1* and *lgerca_lisa_2*, we use the ground-truth shot boundaries provided by MPEG-7 data set. For other videos, we employ the video partitioning algorithm in [26] for shot boundary detection. For each shot, ground-truth objects are manually identified and labelled at every frame. Each object is hand-labeled with a bounding box that minimally encloses its size. All objects in the videos are labeled, except for the objects that are not in motion throughout the shots or their sizes are too small (approximately less than 15×15 pixels) to be significant.

A. Candidate Detection

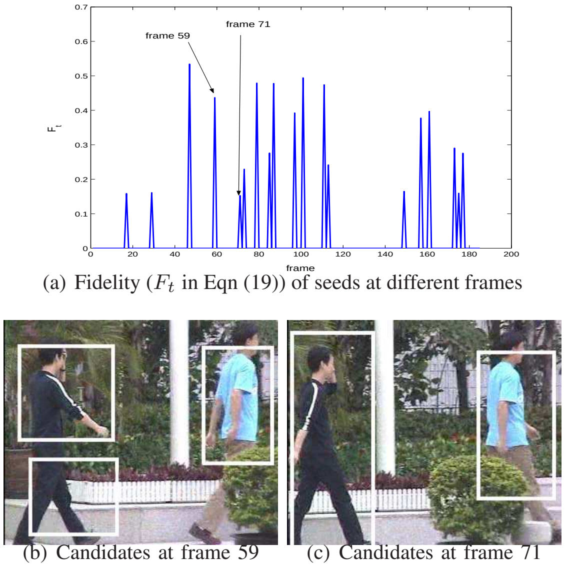
In each shot of videos, a bag of seed candidates is detected based on tensor representation and motion clustering presented in Section IV-A and Section IV-B. Basically each object is associated with a set of candidates, thus the number of candidates is several times larger than ground-truth objects. Table I shows the details. Totally there are 1150 different ground-truth objects being labeled and 20189 candidates being detected. The total number of detected candidates is obtained by counting the number of detected candidates in every frame. In other words, this value is equal to the number of detected objects times the number of frames in which they appear. We manually check the detected candidates and decide the number of correct detection. To evaluate the overall performance, two criteria, CP (candidate precision) and OR (object recall), are used. CP assesses the precision of seed candidate detection. OR measures the capability of recalling ground-truth objects, given the bag of candidates. CP and OR are defined as

$$\begin{aligned}\text{CP} &= \frac{\text{Number of correctly detected candidates}}{\text{Number of detected candidates}}\\ \text{OR} &= \frac{\text{Number of correctly detected objects}}{\text{Number of ground-truth objects}}\end{aligned}$$

Let $\mathcal{G}$ and $\mathcal{D}$ are respectively a manually and automatically detected candidate. A candidate $\mathcal{D}$ is determined as correct detection if there is an overlap of at least 50% in size between $\mathcal{G}$ and $\mathcal{D}$, i.e.,

$$\frac{area(\mathcal{D}) \cap area(\mathcal{G})}{\max(area(\mathcal{D}), area(\mathcal{G}))} \geq 50\% \tag{23}$$

The precise detection of seed candidate is not so critically important at this stage, since EM algorithm and seed association will further refine and update the object boundary. Table I shows the experimental results of candidate detection. On average, the approach achieves approximately 89% of detection precision, while 88% of ground-truth objects are successfully found in the bag of candidates. False alarms do happen mainly due to severe shaking artifact, motion blur and poor visual quality. These false alarms usually appear in few adjacent frames and thus can be readily pruned. The tensor computation and motion clustering methods presented in Section IV actually perform satisfactorily in our data set except for cases like motion blur and illumination change.



(a) Fidelity (F_t in Eqn (19)) of seeds at different frames

(b) Candidates at frame 59 (c) Candidates at frame 71

Fig. 6. The effect of seed candidate selection.

B. Selection and Association

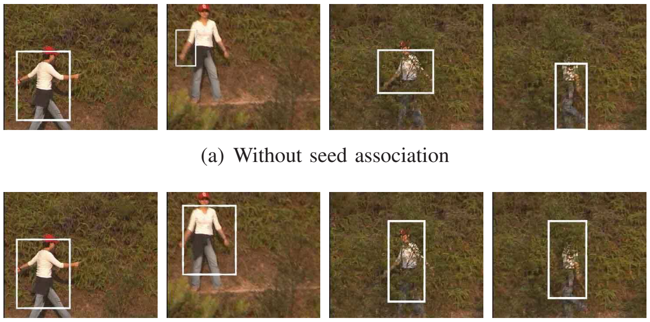
Once seed candidates are detected, best seeds are sequentially selected for subsequent data association. To demonstrate the effectiveness of temporal seed selection, Figure 6 depicts the upshot of selecting a seed from a sequence. Figure 6(a) shows the saliency values (F_t in Eqn (19)) of frames in a shot of *Campus1*. In principle, these values indicate the quality of seeds at different frame, and our aim is to select a seed which can capture object integrity. Figures 6(b)-6(c) show two examples of seeds. By our approach (Eqn (20)), two objects which are integrally detected at frame 71 and with the lowest F_t value are selected as two initial seeds. Obviously, this choice is better than the one in frame 59 where one of the objects is falsely regarded as two seeds. In the experiments, we manually browse all the initially selected seeds in the tested videos, and confirm that this strategy works satisfactory and is able to pick up seeds with most objects appeared in their entirety.

Figure 7 shows an example to illustrate the effect of seed association. In (a), without association, candidates appear in varying forms depending on the underlying motion and scene complexity. By selecting the best candidate (in left image) as seed, association is realized by uniforming the predicted state and current observation (candidate). As noticed in (b), the object is captured nicely after seed association regardless of intermittent motion and partial occlusion which cause serve degradation to observation during the stage of candidate detection. Figure 8 further shows an example of multiple seed selection and association. In this shot, the objects interact with and occlude each other frequently. Our approach successfully selects and initializes three seeds in their entirety in three different frames. These seeds are simultaneously

TABLE I
RESULTS OF CANDIDATE DETECTION

Video	Shot	Time (min)	Ground-truth object	Correct candidate	False candidate	Correct object	Precision (CP)	Recall (OR)
lgerca_lisa_1	42	18	107	506	39	76	0.93	0.71
lgerca_lisa_2	46	18	133	972	58	122	0.94	0.92
hiking	14	16	12	1045	243	12	0.81	1.00
campus1	39	26	111	2186	424	101	0.84	0.91
campus2	37	14	76	2188	409	72	0.84	0.95
harbor1	28	16	97	930	181	80	0.84	0.82
harbor2	30	22	68	2438	258	61	0.90	0.90
mall	11	3	43	329	30	34	0.92	0.79
congregation	16	7	51	673	48	48	0.93	0.94
street	101	34	385	5660	628	341	0.90	0.89
park	17	12	67	862	82	57	0.91	0.85
Total	381	186	1150	17789	2400	1004	-	-
Average	34.6	16.9	104.5	1617.2	218.2	91.3	0.89	0.88

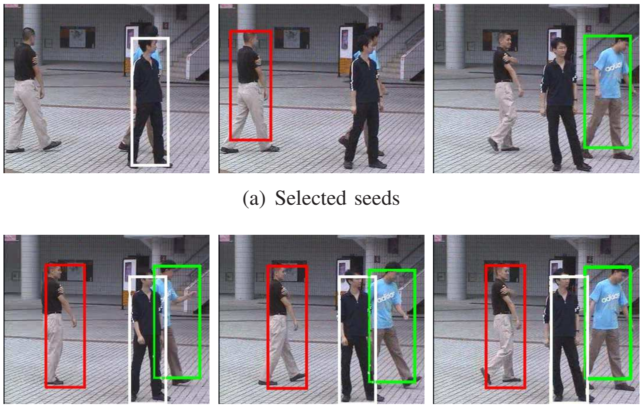
tracked in bi-directional manner throughout the shot as shown in Figure 8(b).



(a) Without seed association

(b) With Seed association (object in left image is selected as initial seed)

Fig. 7. Effect of seed association.



(a) Selected seeds

(b) Associated seeds

Fig. 8. Seed selection and association for multiple objects.

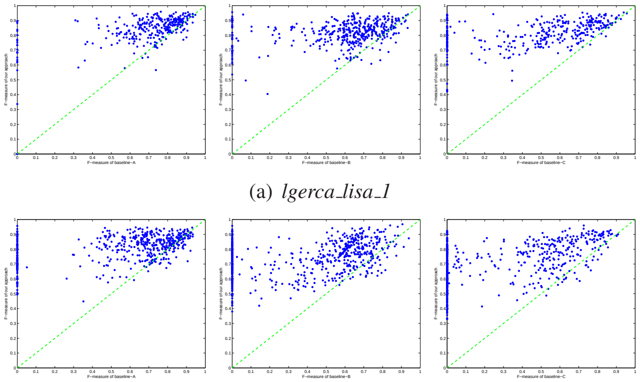
C. Object Initialization and Performance Comparison

Object initialization is fulfilled by combining detection, selection and association. A critical step in this process is the selection of best possible seeds for association, given the bag of candidates with diversity of detection quality. To verify the performance of this step, we divide the bag of candidates in every shot into three bags (B_1, B_2, B_3) according to the goodness of candidate quality based on the measure given in Eqn (19). B_1 contains the best 25% of candidates, B_3 contains the worst 25% of candidates, while B_2 collects the remaining candidates. To this end, we compare our approach with three baseline approaches:

- Baseline-A: Randomly select a seed from the bag B_1 for object initiation.
- Baseline-B: Randomly select a seed from the bag B_2 for object initiation.
- Baseline-C: Randomly select a seed from the bag B_3 for object initiation.

Each bag is basically composed of a subset of seed candidates in a shot. For each baseline, a seed is picked from the corresponding bag as the first seed and then data association is activated to update the seed at every frame of the shot. For our approach, similarly, one starting seed is selected. Based on the sequential selection, more seeds may subsequently be selected depending on the outcome of seed association. In spite of the fact that the first selected seeds of various approaches always start at different frames, the temporal support, more specifically the seed candidates, used for data association at each frame is the same. In other words, the results of the four tested approaches vary in the sense that the information carried and fused from the first selected seed is different and thus affects the performance of object localization at a frame. We utilize this fact to verify that a good quality seed increases the chance of getting better association in home video domain. Thus, in the experiment, the associated seeds (of different approaches) are compared to the ground-truth seeds in a frame-by-frame basis (for frames where seeds appear). Then, the F-measure in Eqn (24) is used to evaluate the performance of the four compared approaches.

The experiment is evaluated based on the 366 shots of the eleven testing videos. For every baseline, one seed per shot is randomly picked from its bag and then temporally associated to locate the object throughout the shot. The quality of located objects at a frame is then assessed by F-measure, precision and recall. Denote $\mathcal{G}$ and $\mathcal{O}$ as the ground-truth and associated objects respectively at current frame, F-measure calculates the fitness of $\mathcal{G}$ and $\mathcal{O}$ by considering the precision and recall of



(a) *lgerca_lisa_1*

(b) *lgerca_lisa_2*

Fig. 9. Comparison to baselines A (left), B (middle) and C (right).

initialization. F-measure is defined as

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times area(\mathcal{G} \cap \mathcal{O})}{area(\mathcal{G}) + area(\mathcal{O})} \tag{24}$$

where

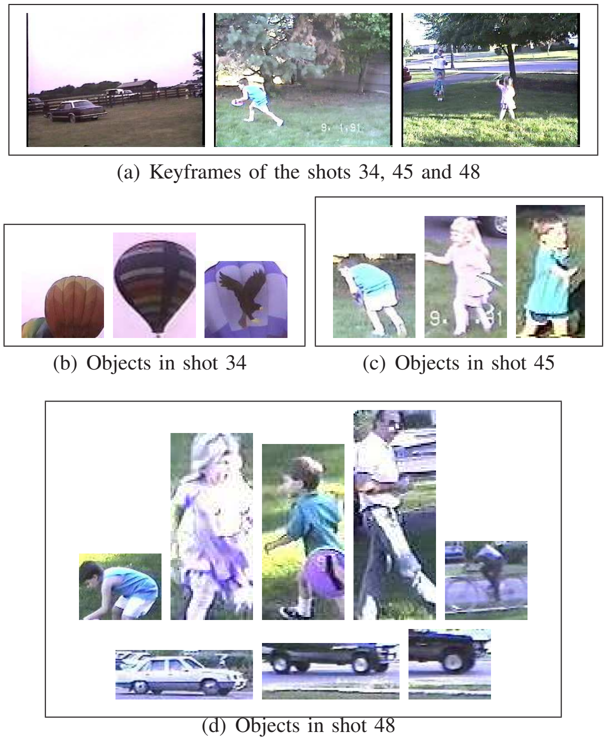
$$\begin{aligned} \text{Precision} &= \frac{area(\mathcal{G} \cap \mathcal{O})}{area(\mathcal{O})} \\ \text{Recall} &= \frac{area(\mathcal{G} \cap \mathcal{O})}{area(\mathcal{G})} \end{aligned}$$

Note that the evaluation is assessed by comparing the associated seeds to the ground-truth objects manually labeled at frame-level. For each shot, the result of an approach is obtained by averaging the F-measure, recall and precision of the associated seeds at every frame.

Figure 9 shows the performance for *lgerca_lisa_1* and *lgerca_lisa_2* based on F-measure. The y-axis denotes the F-measure of our approach, while the x-axis shows the F-measure of three baseline approaches. A point in the plot represents the F-measure of a seed by our approach against one of the baselines. The dotted line marks the border when the performance of two compared approaches is equal. Basically, our approach outperforms others if more points lie above the dotted line, and vice versa. For ease of illustration, we only plot the values of F-measure once per 10 frames. The results indicate that the distribution of F-measure values by our approach is relatively stable where most points concentrate in the range of $[0.7, 1.0]$. In contrast, the distributions of baselines A to C appear diverse and the values of F-measure highly depend on the qualities of seed state and current observation at hand. Also note that many points indeed lie on the y-axes which mean a tracked seed cannot proceed any further due to high uncertainty as well as low confidence of data. This situation happens owing to the fact that the results of association degrade rapidly and accumulate over time as a consequence of inheriting a bad seed for prediction and filtering. Table II summarizes the performance of various approaches in term of F-measure, precision and recall averaged over all seeds in 11 videos. The F-measure of the proposed approach is 0.78, followed by Baseline-A (0.53), Baseline-B (0.46) and Baseline-C (0.39). Overall, our approach performs satisfactorily and outperforms three other baselines by possessing the capability of selecting the best possible seeds as appropriate and then activating association.

D. Video Object Representation

We manually browse through the objects detected in the videos, and find that most objects are indeed correctly initialized in their entirety despite the difficulties arising in home videos. Figures 10 and 11 show the snapshots of some semantic objects detected in *lgerca_lisa_1* and *lgerca_lisa_2*. A large variety of objects including rigid and non-rigid patterns are novelly selected at different frames as starting seeds. Most seeds are correctly initialized and then assembled as the key objects of their shots. Figures 10 and 11 show both keyframes and key objects of shots. From the content management point of view, keyframes provide a global view of frame-level scene activities, while key objects have the capacity of furnishing video database management with object-level browsing and indexing[3].



(a) Keyframes of the shots 34, 45 and 48

(b) Objects in shot 34

(c) Objects in shot 45

(d) Objects in shot 48

Fig. 10. Parsed objects in the 34th, 45th and 48th shots of *lgerca_lisa_1*.

VIII. Summary and Conclusions

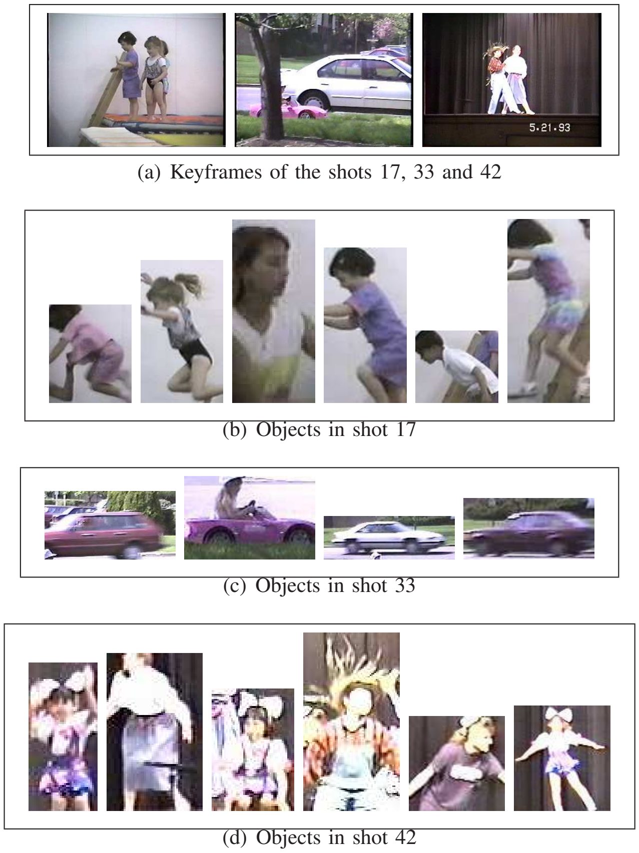
As more and more people possess hand-held camcorders, home video has emerged as a new kind of "multimedia album" for people to document their daily lives with vivid visual content. In this paper, we address the issues of initializing objects in home videos to support object-based parsing and indexing of personal digital video archives. Toward this goal,

[3] Interested readers can find some sample video clips and the extracted video objects by our approach in http://www.cs.cityu.edu.hk/˜cwngo/VOE/VOE.htm. The ground-truth of some videos are available.

TABLE II
PERFORMANCE COMPARISON

Video	Proposed Approach			Baseline-A			Baseline-B			Baseline-C		
	Prec	Rec	FM	Pre	Rec	FM	Prec	Rec	FM	Prec	Rec	FM
lgerca_lisa_1	0.8	0.86	0.82	0.72	0.66	0.68	0.53	0.7	0.59	0.51	0.51	0.49
lgerca_lisa_2	0.71	0.83	0.75	0.56	0.6	0.55	0.4	0.45	0.4	0.26	0.38	0.27
hiking	0.71	0.85	0.76	0.42	0.62	0.46	0.43	0.53	0.45	0.34	0.42	0.36
campus1	0.78	0.86	0.81	0.5	0.66	0.55	0.48	0.57	0.51	0.33	0.42	0.35
campus2	0.76	0.86	0.81	0.54	0.63	0.58	0.42	0.58	0.49	0.44	0.56	0.49
harbor1	0.74	0.78	0.76	0.51	0.43	0.46	0.35	0.45	0.40	0.33	0.37	0.35
harbor2	0.79	0.79	0.79	0.40	0.52	0.45	0.35	0.46	0.40	0.33	0.40	0.37
mall	0.78	0.81	0.79	0.44	0.49	0.46	0.41	0.39	0.40	0.39	0.37	0.38
congregation	0.72	0.81	0.76	0.49	0.52	0.51	0.35	0.51	0.42	0.33	0.30	0.31
street	0.80	0.86	0.83	0.60	0.65	0.63	0.58	0.59	0.59	0.51	0.60	0.55
park	0.63	0.77	0.69	0.36	0.58	0.44	0.30	0.49	0.37	0.30	0.49	0.37
Average	0.75	0.83	0.78	0.50	0.58	0.53	0.42	0.52	0.46	0.37	0.44	0.39

Pre: Precision; Rec: Recall; FM: F-Measure



(a) Keyframes of the shots 17, 33 and 42

(b) Objects in shot 17

(c) Objects in shot 33

(d) Objects in shot 42

Fig. 11. Parsed objects in the 17th, 33th and 42th shots of *lgerca_lisa_2*.

we have presented an offline multiple object initialization approach that considers both the problems of *how* and *when* to initialize objects. We describe several challenges in home video processing and illustrate why deciding when to initialize is an issue for semantic object parsing. Through experiments, we verify that our approach can correctly initialize and track most objects in home videos. In term of speed efficiency, our approach can currently run in 9 frames per second on a 3GHz Pentium-4 machine with 512MB memory.

Although encouraging, the automatic parsing of semantic objects presented in this paper is just a beginning step towards video database management. To further answer whether a detected object is significant is to look into the frequency of object appearance not just within shots but also across shots. This strategy involves object matching which could be challenging if 3D information is taken into account. Another related issue is the categorization (or matching) of video objects within-shot and across-shot for indexing which we do not consider in this paper. Currently, we maintain and update object appearance over time through seed association. This information can be further exploited for categorization by intelligently selecting the key-poses of each seed for within-shot and across-shot object matching.

APPENDIX I

In Section V-C, we employ meanshift for deterministic prediction. Thus, the dynamic model in Eqn (11) can be written as

$$s_t = s_{t-1} + d_t \tag{25}$$

where d_t is the displacement predicted by meanshift algorithm. The deterministic density $p(s_t|s_{t-1})$ in Eqn (12) can be represented by a delta function $\delta(s_{t-1} - (s_t - d_t))$. Thus we can have

$$\begin{aligned} & p(s_t|C_{t-1}) = \qquad (26) \\ & \int_{s_{t-1}} p(s_t|s_{t-1})p(s_{t-1}|C_{t-1}) = p(s_{t-1} = s_t - d_t|C_{t-1}). \end{aligned}$$

When $t = 0$, $p(s_0|C_0) = p(s_0|c_0) = p(c_t|s_0)$ is Gaussian (assuming $p(s_0)$ is uniform), thus $p(s_{t-1}|C_{t-1}) \sim N(\hat{s}_{t-1}, \hat{P}_{t-1})$. This implies that $p(s_t|C_{t-1})$ follows Gaussian distribution,

$$p(s_t|C_{t-1}) \sim N(\tilde{s}_t, \tilde{P}_t) \tag{27}$$

where

$$\begin{aligned} \tilde{s}_t &= E(s_t|C_{t-1}) = \hat{s}_{t-1} + d_t \\ \tilde{P}_t &= \hat{P}_{t-1} \end{aligned} \tag{28}$$

In addition, by Eqn (14), the measurement model $p(c_t|s_t)$ follows Gaussian distribution,

$$p(c_t|s_t) \sim N(s_t, U_t) \tag{29}$$

Combining Eqn (27) and Eqn (29), $p(c_t|C_{t-1})$ becomes,

$$p(c_t|C_{t-1}) = \int_{s_t} p(c_t|s_t)p(s_t|C_{t-1})$$

which is still a Gaussian distribution,

$$p(c_t|C_{t-1}) \sim N(\tilde{s}_t, \tilde{P}_t + U_t) \tag{30}$$

Rewriting Eqn (10) as

$$p(s_t|C_t) = \frac{p(c_t|s_t)p(s_t|C_{t-1})}{p(c_t|C_{t-1})} \tag{31}$$

and substituting Eqn (27), Eqn (29) and Eqn (30) into Eqn (31), we have

$$\begin{aligned} p(s_t|C_t) &= \frac{|\tilde{P}_t + U_t|^{1/2}}{(2\pi)^{n/2}|U_t|^{1/2}|\tilde{P}_t|^{1/2}} \exp\{-1/2[(s_t - \tilde{s}_t)^T \\ & \quad \tilde{P}_t^{-1}(s_t - \tilde{s}_t) + (c_t - s_t)^T U_t^{-1}(c_t - s_t) - \\ & \quad (c_t - \tilde{s}_t)^T(\tilde{P}_t + U_t)^{-1}(c_t - \tilde{s}_t)]\} \end{aligned} \tag{32}$$

Now completing squares in the $\{\}$, Eqn (32) is simplified to

$$\begin{aligned} p(s_t|C_t) &= \frac{|\tilde{P}_t + U_t|^{1/2}}{(2\pi)^{n/2}|U_t|^{1/2}|\hat{P}_t|^{1/2}} \\ & \quad \exp\{-1/2[(s_t - \hat{s}_t)^T \hat{P}_t^{-1}(s_t - \hat{s}_t)]\}, \end{aligned} \tag{33}$$

where

$$\hat{s}_t = \tilde{s}_t + \tilde{P}_t(\tilde{P}_t + U_t)^{-1}(c_t - \tilde{s}_t) \tag{34}$$

$$\hat{P}_t^{-1} = \tilde{P}_t^{-1} + U_t^{-1} \tag{35}$$

or equivalently,

$$\begin{aligned} K_t &= \tilde{P}_t(\tilde{P}_t + U_t)^{-1} \\ \hat{s}_t &= \tilde{s}_t + K_t(c_t - \tilde{s}_t) \\ \hat{P}_t &= (I - K_t)\tilde{P}_t \end{aligned} \tag{36}$$

Notice that Eqn (28) and Eqn (36) form a Kalman filter, with meanshift algorithm as the deterministic dynamic model.

References

[1] Adobe. Premiere. [Online]. Available: http://www.adobe.com/products/premiere.

[2] Apple. imovie. [Online]. Available: http://www.apple.com/imovie.

[3] D. Gatica-Perez, A. Loui, and M.-T. Sun, "Finding structure in home videos by probabilistic hierarchical clustering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 6, pp. 539–548, June 2003.

[4] D. Gatica-Perez and et. al., "Assessing scene structuring in consumer videos," in *Int. Conf. on Image and Video Retrieval*, 2004.

[5] G. Iyengar and A. Lippman, "Content-based browsing and edition of unstructured video," in *Int. Conf. on Multimedia and Expo*, 2000.

[6] W. Y. Ma and H. J. Zhang, "An indexing and browsing system for home video," in *European Signal Processing Conf.*, 2000.

[7] Z. Pan and C.-W. Ngo, "Structuring home video by snippet detection and pattern parsing," in *ACM SIGMM international workshop on Multimedia information retrieval*, 2004, pp. 69–76.

[8] J. R. Kender and B. L. Yeo, "On the structure and analysis of home videos," in *Asian Conf. on Computer Vision*, 2000.

[9] R. Lienhart, "Abstracting home video automatically," in *ACM Multimedia Conf.*, 1999, pp. 37–41.

[10] ——, "Dynamic video summarization of home video," in *SPIE: Storage and Retrieval for Media Database*, 2000.

[11] X. S. Hua, L. Lu, and H. J. Zhang, "Ave - automated home video editing," in *ACM Multimedia Conf.*, 2003.

[12] D. Gatica-Perez and M. T. Sun, "Linking objects in videos by importance sampling," in *Int. Conf. on Multimedia and Expo*, 2002.

[13] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Int. Conf. on Computer Vision*, 2003.

[14] ——, "Video data mining using configuration of viewpoint invariant regions," in *Int. Conf. on Computer Vision and Pattern Recognition*, 2004.

[15] J. Y. A. Wang and E. H. Adelson, "Representation moving images with layers," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 625–638, 1994.

[16] H. S. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 8, pp. 814–830, Aug. 1996.

[17] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *Intl. Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[18] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, 2003.

[19] Y. Weiss and E. H. Adelson, "A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models," in *Int. Conf. on Computer Vision and Pattern Recognition*, 1996, pp. 321–326.

[20] M. A. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[21] A. Mittal and L. S. Davis, "M_2 tracker: A multi-view approach to segmentation and tracking people in a cluttered scene," *Intl. Journal of Computer Vision*, vol. 51, no. 3, pp. 189–203, 2003.

[22] M. Yeasin, E. Polat, and R. Sharma, "A multiobject tracking framework for interactive multimedia applications," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 398–405, 2004.

[23] U. Neumann and S. You, "Integration of region tracking and optical flow for image motion estimation," Oct 1998, pp. 658–662.

[24] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. Wiley New York, 1987.

[25] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[26] C.-W. Ngo, T.-C. Pong, and R. T. Chin, "Video partitioning by temporal slice coherency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 8, pp. 941– 953, 2001.

PLACE
PHOTO
HERE

Zailiang Pan received his MPhil in Computer Science from the City University of Hong Kong in 2005. He received his BSc in Computer Science from Tsinghua University of China in 2002. His research interests include multimedia information retrieval, computer vision, pattern recognition and machine learning.

PLACE
PHOTO
HERE

Chong-Wah Ngo (M'02) received his Ph.D in Computer Science from the Hong Kong University of Science & Technology (HKUST) in 2000. He received his MSc and BSc, both in Computer Engineering, from Nanyang Technological University of Singapore in 1996 and 1994 respectively.

Before joining City University of Hong Kong as assistant professor in Computer Science department in 2002, he was a postdoctoral scholar in Beckman Institute of University of Illinois in Urbana-Champaign (UIUC). He was also a visiting researcher of Microsoft Research Asia in 2002. His research interests include video computing, multimedia information retrieval, data mining and pattern recognition.