Multi-view Video Based Object Segmentation - A Tutorial

ChunHui Cui[1], **Qian Zhang**[2], and **KingNgi Ngan**[3], Non-members

ABSTRACT

Video based object segmentation (VBOS) is an important step in many computer vision and multimedia tasks such as video editing and compositing. In recent years, multi-view VBOS systems have become more and more popular because the stereo clues from multi-view data can be efficiently incorporated to improve the segmentation results and eliminate the required initial user input. In this paper, we give a review on recent development of multi-view VBOS systems and the related techniques including data acquisition, camera calibration, depth reconstruction, object segmentation and tracking. Furthermore, we introduce our multiple objects segmentation system from multi-view video sequence to illustrate the practical implementation of multi-view VBOS system for 3D video rendering applications.

Keywords: Multi-view video data acquisition, camera calibration, depth reconstruction, object segmentation and tracking

1. INTRODUCTION

Video segmentation is a crucial technique in many video applications such as object-based video coding, video surveillance and video conferencing. Generally, the approaches can be categorized into two classes: shot-based segmentation and object-based segmentation. Shot-based segmentation partitions the video sequence into a set of video shots and extracts key-frames to represent each video shot. In comparison, object-based segmentation focuses on decomposing the video shot into objects and background. Based on different semantic level of representing a video shot, these two segmentation approaches are usually implemented independently or jointly for a variety of advanced video analysis tasks.

The development of MPEG-4, MPEG-7 and many supported applications intensify the development of video-based object segmentation (VBOS) systems. Basically, the scene can be decomposed by segmentation into a few meaningful objects that are represented as distinct spatial-temporal entities. Thus, object-based representation can be achieved by segmenting each frame, and visual information can be coded, viewed and edited once these segmentations are available. To achieve accurate segmentation results, user interaction is usually involved in many advanced VBOS systems. Wang et al. [1] present an interactive system for efficient segmentation of foreground objects from a video sequence. It mainly consists of three stages: automatic preprocessing, interactive segmentation and post-processing. Employing automatic preprocessing, a hierarchical decomposition of the input video is pre-computed by interactively applying a mean-shift clustering algorithm. In the interactive segmentation stage, a novel user interface is built, which allows users to indicate the foreground object across space and time. Then a novel version of min-cut optimization over the entire video is proposed to compute a coarse segmentation based on the hierarchical representation. Finally, in order to refine the object boundary, the post-processing stage performs a pixel-based min-cut optimization constrained to the narrow region around the segmentation results in the previous stage. A spatio-temporally coherent alpha matte enables the foreground object to seamlessly merge onto other background. Another interactive video-based object segmentation system is introduced in [2], where the object cutout can be pasted onto another sequence or image. Firstly, the users are required to select a few key frames in the video sequence. Then a novel 3D graph-cut based segmentation is performed between each pair of successive key frames to extract the accurate silhouettes of multiple objects. For accuracy, the system allows users to refine the segmentation results in the video tube extracted by the bi-directional featured tracking algorithm. Finally, the video object is cutout by coherent matting within the trimap and then object paste can be achieved.

Recently, due to the growing availability of inexpensive video cameras and new generations of more powerful computers, multi-view segmentation techniques have attracted an increasing interest from both the research community and industry. By incorporating the additional depth information, robust segmentation results can be achieved and the need for user interaction can be significantly reduced as well. Cardoso et al. [3] investigate the issue of automatic object-based spatial video segmentation assisted by depth and motion information. They started at the suggestion of a practical framework that incorporates the depth information into the segmentation process

Manuscript received on July 10, 2008.

[1,2,3] The authors are with the Department of Electronic Engineering, The Chinese University of Hong Kong, HongKong Tel.(+852) 2609 8255 , Fax:(+852) 2603 5558, Email: chcui, qzhang, knngan @ee.cuhk.edu.hk.

of a color image, where the depth information is used to guide image segmentation by automatically estimating the number and localization of objects. Then the motion information is involved to improve the segmentation quality of moving object in dynamic sequence. Compared with the single object segmentation, multiple objects segmentation is a more general and difficult issue. Reid et al.[4] proposed an algorithm for multiple objects segmentation from the multi-view video of a dynamic scene. This approach uses maximum a posterior (MAP) estimation to compute the parameters of a layered representation of the scene, where each layer is modeled by its motion, appearance and occupancy. The MAP estimation of all layer parameters is equivalent to tracking multiple objects in a given number of views. Expectation-Maximization (EM) is employed to establish the layer occupancy and visibility posterior probabilities. The experiments on single-view and two-view video data of *Football* sequence show the efficiency of segmenting and tracking non-rigid objects even with extreme occlusion.

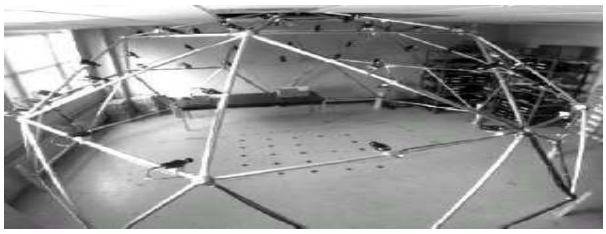
This paper mainly focuses on introducing the newly emerging multi-view VBOS systems and the related techniques such as data acquisition, depth reconstruction, object segmentation and tracking. The remainder of the paper will be organized as follows. Section 2 introduces a variety of multi-view video capturing systems and a number of representative camera calibration approaches. Section 3 gives an overview of stereo matching algorithms for depth reconstruction. In Section 4, object segmentation algorithms using graph cut will be investigated. Section 5 describes the development of our multi-view VBOS system as an illustration of practical implementation. Finally, the conclusion will be drawn in Section 6.

2. CAPTURING SYSTEM FOR MULTI-VIEW DATA

2.1 Multi-view camera system

Data acquisition is the first step but a very important component in a VBOS system. By capturing high quality images, we can reduce a lot of artifacts and avoid complicated post-processing. In order to build a suitable video capturing system, two factors should be taken into consideration, i.e., choosing the right type of cameras and constructing a proper camera configuration. In addition, the associated control unit and processing software should be employed to work with the capture system. In recent years, with the fast improvement of the capability of personal computers and digital equipments, more and more multi-camera systems with dense or sparse, wide-baseline or narrow-baseline camera configuration become available, which significantly broaden the application areas and enhance the user experience.

The Virtualized Reality Project, started in 1996 and developed by the CMU Robotics Institution, is considered to be the pioneering work in this field. CMU *3D-Dome* [5] is probably the first integration of synchronized cameras at multiple viewpoints. As shown in Fig 1(a), the *3D-Dome* consists of 51 cameras mounted on a 5-meter diameter geodesic dome. It uses consumer VCRs to record the synchronized video from each monochrome CCD camera. The resolution of each camera is 512×512 and the capture rate is 30 frames per second (fps). The camera configuration shown in Fig 1(b) is the second generation, CMU *3D-Room* [6]. The cameras are arranged in a $6.1(L) \times 6.1(W) \times 2.7(H)$ meters room, 10 on each of the 4 walls and 9 on the ceiling. It makes use of 49 synchronized color S-Video cameras to capture the 640×480 video at 30 fps. The computing system of the *3D-Room* is composed of a control PC and 17 digital PCs. By now, this project has been developed to the third generation, *3D-Cage*. It has 48 cameras controlled by 25 PCs mounted on the gird of all walls in a room showed in Fig 1(c). It can continuously capture full color images with 640×480 resolution at 30 fps for over 2 hours



(a)3D-Doom

(b)3D-Room

(c)3D-Cage

Fig.1: *Virtualized Reality Project in CMU.*

A self-reconfigurable camera array system that can interactively capture and render 3D virtual scene is presented in [7]. This camera array in Fig 2 is composed of $48(8 \times 6)$ *Axis 205* network cameras located on 6 linear guides. It can capture up to $640 \times 480 \times 30$ videos. The cameras have built-in HTTP servers to respond to HTTP request and send out motion JEPG sequences. The most distinguishing characteristic of the system is its reconfiguration because the cameras are mounted on a mobile platform.

Fig.2: *A self-reconfigurable camera array.*

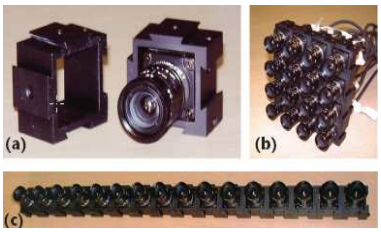


Fig.3: *Camera Array Pursuits for Plenoptic Acquisition.*

In [8], a versatile camera array system named *Camera Array Pursuits for Plenoptic Acquisition* (CAPPA) is constructed, with 16 NTSC color camera heads of *Sony XC-333*. This camera system can be arranged in different configurations with modular camera head units in Fig 3(a), densely arranged in a lattice as in Fig 3(b), or connected in a row as in Fig 3(c). Moreover, it can construct a sparse camera array by inserting empty units between camera head units. The lenses of the individual cameras can be altered to capture light rays from several viewing angles. The distance between neighboring cameras is about 31mm and can be adjusted to 62mm in the sparse configuration. In order to achieve real-time processing, the camera system is arranged as shown in Fig 3(b) in practical implementation. The quad processor (*Sony YS-Q430*) combines the videos from these 16 cameras into a single 16-screen sequence, followed by an *SGI Onyx2* workstation with a DIVO (digital video option) to capture and process this single 16-screen sequence

Apart from these large camera arrays with tens of cameras, some multi-view capturing systems adopt sparse configuration with small number of cameras to capture the images for certain applications. The multi-camera system [9] developed by MSR uses an 8-camera array to capture the videos off-line. Eight cameras are placed along a 1D arc spanning about from one end to the other as shown in Fig 4. The images are captured with high resolution 1024×768 at 15 fps by high-quality *PtGrey* color cameras. The synchronized video stream from multiple views is combined to accomplish the real-time and high-quality dynamic scene rendering with interactive viewpoint control. Intermediate view generation of *Soccer* scene is the objective in [10]. In this system, a group of four cameras is placed to one side of the field to capture the penalty area as shown in Fig 5. In [11], the impressive input data captured by a four-camera array is utilized to investigate an image-based approach for computing and shading visual hull. In addition, five calibrated cameras are used for on-line 3D scene acquisition and view synthesis [12].

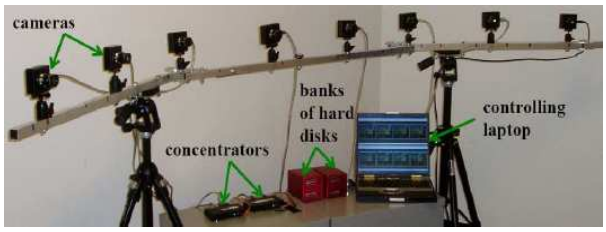


Fig.4: *MSR camera array.*

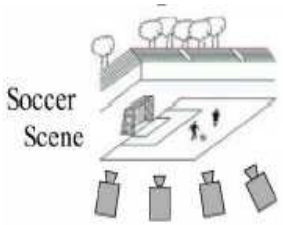


Fig.5: *4-camera array capturing soccer scene.*

3DTV and FTV are the new forms of media communication representing very important multi-view applications. The multi-view image capturing system for FTV application in [13] is shown in Fig 6. The acquisition system consists of one host-server PC and 100 client PCs. Each client PC associates with a high-definition camera (*JAI PULNiX TM-1400CL*). The system is able to capture 100 synchronized high-resolution video signals at 30 fps. Additionally, the camera position can be tuned with ease. The acquisition system generates high-quality free viewpoint images at different times and positions.

2.2 Multi-camera calibration

As introduced in section 2.1, a typical multi-view capturing system usually requires multiple cameras to capture scenes of considerable extent in large rooms or even outdoors. A complete multi-camera calibration is an inevitable and important step towards the efficient use of such systems. On the one hand, the epipolar geometry recovered by calibration significantly reduces the search space of disparity, and also helps to eliminate many erroneous correspondences due to similar local appearance. More importantly, camera calibration enables one to obtain the geometry of perspective projection, which is the basis for depth reconstruction and novel view rendering. In

Fig.6: *Multi-view capturing system for FTV.*

this section, we start with the basic notation in camera calibration and then briefly discuss a variety of multi-camera calibration methods.

2.2.1 Basic notation

The projective geometry of the pinhole camera is modeled by perspective projection. A 3D point, $M = [x, y, z]^T$, whose coordinates are expressed with respect to (w.r.t.) the world coordinate system, will be projected onto the image plane with coordinates $m = [u, v]^T$. Their homogeneous coordinates are denoted by $\overline{M} = [x, y, z, 1]^T$ and $\overline{m} = [u, v, 1]^T$. The relationship between a 3D point M and its image projection m is given by

$$M \simeq Pm \tag{1}$$

where $\simeq$ indicates equal up to scale and P is the projection matrix determined by the camera intrinsic and extrinsic parameters as

$$P = A[R\ t] \tag{2}$$

The extrinsic parameter (R, t) is the rotation and translation that relates the world coordinate system to the camera coordinate system. The intrinsic matrix A determines how the image coordinates of a point are derived, given the spatial position of the point with respect to the camera. It is given by

$$A = \begin{bmatrix} \alpha f & \gamma & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where (u_0, v_0) is the coordinates of the principal point, f is the focal length, α denotes the aspect ratio and γ is the parameter describing the skewness of two image axes.

2.2.2 Camera calibration methods

The task of multi-camera calibration is to estimate the intrinsic matrix A of each camera and recover the geometrical relationship (the rotation R and translation t) between the camera and the scene, or between different cameras. In addition, the distortion coefficients, mainly representing the radial and tangential lens distortion, will also be estimated in camera calibration. Considerable research has been carried out on camera calibration in the past decades [14] [15].

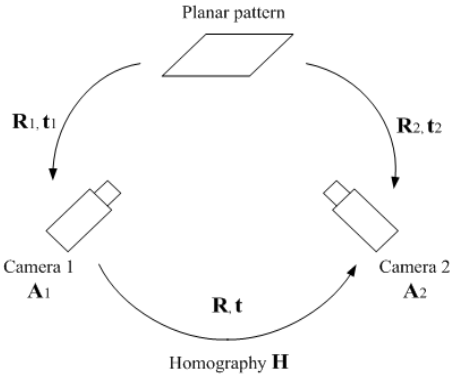


Fig.7: *Homography between two views.*

Here we put more emphasis on a number of representative algorithms.

One of the most popular and accurate calibration methods is proposed by Tsai [16]. Tsai's method represents the conventional world-reference based approach [15] [17] [18] [19] that relies on accurate 3D coordinate measurement with respect to a fixed reference. Camera calibration is performed by observing a calibration object whose geometry in 3-D space is known with very good precision. Even though expensive calibration apparatus and elaborate setup are required, these approaches have been widely used in multi-camera systems due to their accurate calibration results. In contrast, auto-calibration [20] [21] does not use any calibration object. By moving a camera in the static scene, the rigidity of the scene will provide constraints on the intrinsic parameters from camera displacement. Therefore, both the intrinsic and extrinsic parameters can be recovered given a few feature correspondences in multiple scene views. Though this approach is very flexible, reliable results cannot always be obtained since there are many parameters to be estimated [22].

Zhang's [23]method represents the newly-developed planar auto-calibration approach [24] [25] that combines the advantages of both world-reference based and auto-calibration approaches. The basic idea is to use relative geometric information of planar feature points, i.e. homography between model planes. The planar auto-calibration method is flexible in that either the camera or the planar pattern can be moved freely and the calibration results are easily repeatable without redoing any measuring tasks. Moreover, because the planes provide reliable features (pattern corners) and allow accurate intensity-based matches, this approach can achieve high precision that is comparable to or even better than many world-reference approaches. Since many researchers rely on Zhang' method to estimate the camera's intrinsic parameters because of its efficiency, their focus is diverted to the external camera calibration. In the multi-view system [26], positions and orientations of the model planes relative to the camera are first estimated by Zhang's

method. Using this information, rigid transforms between two cameras are then determined through an arbitrarily chosen plane. Besides, a RANSAC procedure is applied to remove possible outliers. The method proposed in [27] estimates the pair-wise relationship based on the epipolar geometry. Translation and rotation between two cameras can be recovered by decomposing the associated essential matrix as the camera intrinsic matrices are obtained beforehand by planar auto-calibration.

Other existing techniques exhibit a few desirable features, although the precision of their calibration results is not comparable to Tsai's and Zhang's method. Calibration method in [28] uses geometric invariants, e.g. parallel lines and vanishing points, and may require special equipment for measuring certain variables. In order to calibrate multiple cameras simultaneously, a few methods [29] [30] [31] have been developed based on matrix factorization and global constraints. Usually the whole projection matrix P is estimated instead of distinguishing intrinsic and extrinsic parameters.

The world-reference based method involves the design and use of highly accurate tailor-made calibration patterns, which is often difficult and expensive to manufacture. The planar auto-calibration approach uses a simple planar pattern instead. However, it will be tedious and annoying to use this approach for calibrating a large number of cameras distributed widely, since points on the planar pattern may not be simultaneously visible in all views. Therefore, recent work on multi-camera calibration tends to use some common simple object such as spheres, to replace the calibration pattern. Spheres were firstly used to compute part of the calibration parameters [32] [33]. Recently, Teramoto et al. and Zhang et al. [34] [35] have successfully extended the method to calibrate the full set of parameters by relating the absolute conic with the images of spheres. The spheres are cheap and ubiquitous calibration object. Their silhouettes can be extracted reliably from images and this facilitates precise camera calibration. Besides, as long as the sphere is placed in the common field of views, its contours are always visible from any viewpoints. Hence spheres can be used to calibrate multiple cameras mounted at arbitrary locations. Other approaches [31] [36] using a laser pointer or virtual calibration object are also suitable for multi-camera calibration. But these methods either involve elaborate feature tracking or have some particular requirements in the scene captured, which limit their applications.

3. DEPTH RECONSTRUCTION

Depth map of the 3D scene is a useful and important cue for multi-view analysis. Generally, the depth value of a 3D point can be reconstructed from multiple disparity maps of multi-view image set. Thus accurate stereo matching becomes crucial issue for the reconstruction of high-quality depth map. Stereo matching is a fundamental problem in computer vision and has been extensively researched in recent years. Given two or more images of the same scene taken from different viewpoints, the goal of stereo matching algorithm is to find the corresponding points in other images for a pixel in one image. The point correspondences between two images generate the disparity map that indicates the difference in locating corresponding points. Based on the perspective projection model, depth can then be reconstructed by disparity map and the calibrated camera parameters.

There are many issues that make stereo matching an unresolved and challenging problem. First, due to the limitations in imaging techniques and the imperfect illumination conditions, raw images may not conform to the photo-consistency rule. In other words, different projections of the same 3D point may have quite different intensity/color values. Second, stereo matching is inherently an ill-posed problem. The textureless regions lack image information for reliable matching, and the occluded regions will have no correspondences. All these ambiguities may lead to incorrect matches. Hence various constraints derived from the inherent properties of disparity are proposed to solve this problem. Moreover, with middle to wide baseline distance in practical multi-view systems, disparity exists in both horizontal and vertical directions, and the search range of disparity values can be very large, which makes disparity estimation even more time-consuming.

A vast number of publications have been dedicated to the stereo matching problem and recently many advanced techniques emerged. A comprehensive survey on stereo matching algorithms can be found in [37]. In general the taxonomy of stereo matching algorithms is based on two categorization criteria. The first one, namely the representation, refers to how to choose the disparity primitive from the input images. According to the scene representation, various stereo matching approaches can be classified as pixel-based and region-based. The pixel-based representation is the most universal one and can be applied to any scene, while the region-based stereo algorithms [38]-[47] have inherent ability of occlusion handling around region boundaries and the better performance in ambiguity discrimination. The second criterion is from the view of the optimization techniques used in disparity estimation. Accordingly the stereo matching algorithms can be classified into local and global methods. Local approaches find the correspondence of each image point individually by examining the similarity of color or intensity values. In contrast, global approaches determine all the disparities simultaneously by applying energy minimization techniques such as graph cuts [48] [49] [50], belief

propagation [51] [52] [53] and dynamic programming [54]-[61]. According to recent advances [62], region-based stereo methods are more favored due to their better disparity smoothness regularization. As for optimization techniques, global approaches in general produce more accurate disparity map, while local approaches have slightly poor results, but is superior with respect to computational complexity.

3.1 Region-based stereo

Region-based stereo matching algorithms assume that all the pixels in a support region have similar disparities. Their performance greatly depends on how the support region is selected for each image point. If the support region actually contains depth discontinuity, the description of the local appearance will be inaccurate, resulting in blurred object boundaries and the removal of details. A number of methods have been proposed to solve the problem, among which adaptive window approach and segmentation-based approach are representative [38]-[41].

Adaptive window algorithms try to find the optimal windows by adaptively changing the window size and shape. The idea of adaptive window can be traced back to Kanade and Okutomi [38], where the appropriate window for each pixel is selected by examining the local intensity and disparity variations. However, this method is computationally expensive and hence the window shape has to be constrained as a rectangle. Veksler [39] presents an algorithm to select a certain window size and shape from a large class of predefined windows. Efficient optimization over many windows is achieved using the minimum ratio cycle algorithm for graphs. Yet the window shape is still not general and the computation of window cost requires many parameters. Tang [40] presents a method of region growing with an adaptive window, but the weight of the window only depends on the geometric distance. Yoon and Kweon [41] propose a general method to determine the optimal local support window. A fixed sized support window is used, but the weight varies for each pixel within the window based on color similarity and geometric distance to the center pixel of interest. Good results can be achieved even without global optimization technique, though it is very time-consuming to compute the pixel-wise support weight.

Based on the reasonable assumption that neighboring pixels with similar color or intensity have similar or continuous depth, researchers have incorporated image segmentation to simplify the stereo matching problem [42]-[45] and achieved good results. Segmentation based approach can reduce the ambiguity of the textureless regions. The side effect of this assumption is that depth discontinuities tend to occur at color boundaries. However, segmentation itself remains a difficult and unsolved problem, and jointly performing disparity estimation and segmentation dramatically increases the solution space and the complexity. For computational efficiency, Gong [46] proposes an adaptive cost aggregation scheme using edge detection instead of color segmentation for real-time implementation with graphic hardware. Yoon et al. [47] use boundary information to compute accurate windows for each pixel, and solve the iterative problem in a hierarchical framework, known as the scale-variant iterative scheme.

3.2 Global optimization techniques

The local methods emphasize the matching cost definition and cost aggregation steps. The final disparities are computed by simply choosing the disparity at each pixel associated with the minimum cost value, i.e., a local "winner-take-all" (WTA) optimization. In contrast, most global methods are formulated in an energy minimization framework [63]. The objective is to find a disparity function d that minimizes a global energy, where various constraints are applied to reduce the uncertainties of disparity map.

$$E(d) = E_{data}(d) + \lambda E_{smooth}(d) \tag{4}$$

where the data term $E_{data}(d)$ measures how well the disparity function d agrees with the input image pair, and the smoothness term $E_{smooth}(d)$ imposes the piecewise smoothness constraint. Once the global energy has been defined, a variety of algorithms can be used to find a (local) minimum. Compared with many traditional regularization methods such as simulated annealing [64] [65], recent advanced techniques, including graph cut and belief propagation, produce much better results.

In recent years, graph cut [48] [49] [50] and belief propagation [51] [52] [53] have been widely used to solve the stereo matching problem based on the energy minimization framework. These methods are efficient in the sense that both the found local minima are minima over "large neighborhoods", and they produce highly accurate results in practice. A comparison between these two different approaches for stereo matching can be found in [54]. However, both the graph cuts and belief propagation approaches have their limitations in application. They achieve good results for frontal-parallel stereo images with small baseline distance, i.e., when the disparity is one dimensional with small range. When applied to wide baseline multi-view images, the label space is much larger and the computational requirement is intensive yet the results are not encouraging. Efforts have been made to reduce the computational complexity. For example, in [55] some algorithmic techniques are proposed to improve the runtime of the loopy belief propagation approach.

As one of the earliest optimization frameworks introduced for stereo matching, dynamic programming (DP) [56]-[59] is still popular because of its high efficiency in many applications. However, traditional

DP optimization on the 1D scan-line structure suffers from the "streaking" problem. Hence many improvements are made to reformulate the stereo matching problem so that it can apply DP optimization on a 2D structure. In [60], by enforcing piecewise smoothness constraint on the horizontal and vertical directions using a two-pass 1D-DP approach, a 2D optimization is simulated. In [61], a more superior reformulation is proposed in which the image is represented by a 2D pixel-tree structure instead of individual 1D scan-lines.

4. GRAPH CUT BASED MULTI-VIEW OBJECT SEGMENTATION

4.1 Background on graph cut theory

Graph cut theory was firstly applied in the field of computer vision at the end of 1980s' in [66]. Greig. et al. showed that certain energy functions formulating some vision problems can be efficiently solved by the powerful min-cut/max-flow algorithm. Image restoration was taken as an example by obtaining the maximum a posterior (MAP) estimation of a binary image using the graph cut technique. However, the potential of the graph cut technique has been unnoticed for ten years because the application of image restoration is limited. Fortunately, with the development of this technique by Boykov and Jolly [67], graph cut has witnessed an extensive interest and turned out to be a popular method for solving many general problems in computer vision, such as image segmentation, stereo matching and multi-camera scene reconstruction.

Graph cut based methods construct a graph topology to minimize the specified energy function activated by the max-flow/min-cut algorithm, so that the min-cut on the graph is of minimal energy among all the cuts separating the terminals. Theoretically, $G = \{V, E\}$ is a directed graph with weighted edge capability. V is the vertices set including all the nodes as well as another two special terminals usually called *source* and *sink*. Generally, nodes represent pixels or voxels, and source and sink correspond to the set of labels assigned to the nodes. E is a directed edge set, which connects all the vertices in the graph. It contains two types of edges in the graph named *T-Link* and *N-Link*. T-Link connects pixels with terminals or labels. N-Link joins the pixels or voxels with their neighbors. All the edges in the graph are assigned weights or costs representing the edge capability. A S-T cut C with two terminals is a partition of nodes in the graph into two disjoint sets S and T so that the source is in S and sink is in T. In combinatorial optimization, the cost of a cut is calculated as the sum of the edge weights passing from S to T. The min-cut problem is to find a cut with the minimal cost. The Ford and Fulkerson theory [68] demonstrates that the min-cut is equivalent to max-flow from S to T which can saturate the capability of the edges and separate

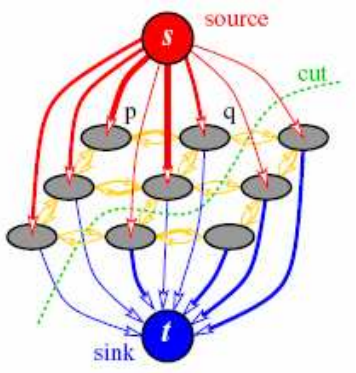


Fig.8: *Graph with a cut.*

the nodes in the graph into two disjoint set $\{S, T\}$. Fig 8 is the graph with a cut.

The min-cut/max-flow algorithm aims to minimize the energy function over the image labeling. How to define the energy function for efficiently representing the properties of the image becomes the key problem in the graph cut theory as well as its applications. For image segmentation, the general formulation of the energy function is defined as follows:

$$E(f) = \sum_{(p \in P)} D_p(f_p) + \lambda \sum_{(p,q \in N)} V_{p,q}(f_p, f_q) \qquad (5)$$

where, f is the labeling field, P is the set of pixels, and N is the pixel's neighborhood system. The first term $D_p(f_p)$ is called the data term and corresponds to the T-link in the graph. It measures the penalty of a certain pixel p assigned to the label f_p and how well the label f_p fits the pixel p given the observed data. The smoothness term $V_{p,q}(f_p, f_q)$ can be represented by the N-link in the graph. It evaluates the penalty of disagreement between p and q which are assigned with f_p and f_q, respectively, and preserve the discontinuity to avoid f smoothing everywhere. λ is a parameter to weigh the importance of these two terms.

Graph cut is not limited to solving the regular binary-label segmentation problem, but is also applicable to multi-label energy minimization. Multi-way cut based on expansion move (α-expansion or α-β-expansion) [69] is proposed to handle the multi-value labeling by repeatedly minimizing the energy function with only binary variables. In multi-label segmentation, $V = \{P, L\}$ where L is the label set with multiple terminals. A subset of edges $C \in E$ is called a multi-way cut if the terminals are separated into induced graph $G(C) = \{V, E - C\}$. The cost of a multi-way cut is equal to the sum of the edge weights removed in the cut. The multi-way cut problem is to find the minimal cost multi-way cut. The expansion move algorithm is used for the standard two-frame stereo vision problem as well as the multi-view scene reconstruction, where the labels correspond to dis-

parities.

4.2 Graph cut based image segmentation

Image segmentation, i.e., the region-partitioning or the pixel-labeling, can be formulated by an energy minimization problem. The graph cut technique works as a powerful tool for energy minimization, and has been successfully applied to solve related vision problems. Object segmentation aims to separate the interested objects from the background. The segmentation results can be employed for background substitution, object-based coding as well as object-based rendering.

The efficiency of graph cut for image segmentation has been demonstrated in many literatures. With the publication of Boykov and Jolly's paper [67], graph cut has become a leading method for N-D image segmentation. This paper describes a technique used for interactive segmentation of N-D images. Users are required to indicate the certain "$foreground$" or "*background*" pixels as hard constraint. Additionally, the soft constraint involving the regional and boundary properties of the segments is incorporated into the energy function. The graph cut is used to find out the globally optimal segmentation of the N-D images. *Lazy Snapping* [70] is an interactive image cutout tool using graph-cut. This tool designs a novel coarse-to-fine user interface (UI), which enables the users to specify the object and background by marking a few lines in the images at the coarse step, and revise the object boundary to achieve refinement at the fine step. Firstly, a pre-computed and over-segmented image is built using watershed algorithm [71] to improve the segmentation efficiency. Then at the coarse step, the graph cut is performed to cut out the object with the help of object marking, where the nodes in the graph are the segmented regions obtained from the watershed algorithm instead of pixels. The formulated energy function encodes the color similarity of nodes in the likelihood term and the color gradient along the object boundary in the prior term. At the fine step, the pixel-based graph cut in a small band with the boundary editing is introduced to refine the ambiguous and low contrast object boundaries.

Because of the efficiency of image segmentation using graph cut, a lot of derivations has emerged based on this technique, such as *GrabCut* [72], *Ratio Cut* [73] and *Normalized Cuts* [74]. GrabCut is an interactive image segmentation technique by iteratively using graph cut. Compared with the graph cut, the user interaction is significantly simplified by drawing a rectangle around the object instead of many lines in the image. The energy function is iteratively minimized using min-cut by learning Gaussian Mixture Model (GMM) parameters. Furthermore, good quality alpha mattes for the foreground can be obtained with modest user effort. Ratio Cut algorithm proposes a new cost function named cut ratio for graph cut based image segmentation. The cut ratio is defined as the ratio of the corresponding sums of two different weights associated along the cut boundary in an undirected graph. Minimization of the cut-ratio cost function is an NP-hard problem that can be solved by a polynomial-time algorithm. Compared with other graph cut based segmentation algorithms on medical and natural images, the results demonstrated that the Ratio Cut is a promising and efficient approach for image segmentation. In Normalized Cuts, image segmentation is regarded as a graph partitioning problem and a novel global criterion is proposed to segment the graph, namely Normalized Cuts. This criterion evaluates both the disagreement between the different groups as well as the similarity within the groups, and can be optimized by efficient computational techniques based on the eigenvalue problem. In addition, this approach is applicable to not only the static images but also the motion sequences.

Most of the classical graph cut based segmentation algorithms ask for users' intervention to provide the initial foreground and background data for modeling. Even though good performance can be achieved, a major drawback is the dependence on these segmentation initializations. First, the interaction itself may be annoying to the users especially it is mandatory. Furthermore, the graph cut suffers from incomplete initialized information. These disadvantages drive the development of intelligent and automatic graph cut based segmentation algorithms. Mu. et al. [75] present a coarse-to-fine algorithm to automatically segment the moving objects from videos. At first, the block-based segmentation is performed on each frame using color and motion information to divide the image into foreground, background and boundary blocks. Then the prior knowledge for data modeling is automatically obtained from foreground and background blocks. Finally, the refined segmentation region is restricted in the boundary block, and graph cut is incorporated to get accurate pixel-wise segmentation result. In [76], the image segmentation is formulated as the region labeling problem. It proposes an iterative optimization scheme to estimate the label configuration by alternately performing MAP estimation and the maximum-likelihood (ML) estimation. The ML estimation is achieved by finding the means of the region features, while the MAP estimation is modeled by Markov random field (MRF) and solved using graph cut to find a solution of MAP-MRF estimation. This algorithm can automatically segment the image into regions with relevant color and texture information without the prior knowledge of the region number.

4.3 Multi-view image segmentation

In many video rendering applications, the end-users desire to render only the interested objects instead of the whole scene. Segmenting out the object from the background is the first and crucial step towards this goal. To provide the capability of all-around viewing of a dynamic object in an object-based video rendering system motivates the research on multi-view object segmentation.

Graph cut based multi-view object segmentation algorithms have been explored in the literature. A background removal algorithm is proposed in [77] to segment every single view of the multi-view images by learning shape priors. The automatic segmentation is initialized by graph cut with trimap labeling. Then, user interaction is involved to choose a subset of the segmentation results with satisfactory quality for 3D reconstruction and learning shape priors. At last, those unsatisfactory results will be refined with the help of the learned shape priors to improve the performance. However, only one interested object can be segmented and the whole procedure is semi-automatic. A 3D-reconstruction and background separation approach for multi-view images is presented in [78], which retrieves the depth map and segments object simultaneously. The energy function is composed of photo-consistency, smoothness, visibility constraints and background properties, so as to encode the high-level knowledge about the scene reconstruction. A strong local minimum can be found using graph cut. Nevertheless, a successful reconstruction and segmentation has to rely on the additional information extracted from a known static background image. In [79], a novel framework to extract the foreground objects from the short-baseline image sequences is proposed for 3D scene reconstruction. The image sequences are first partitioned into a number of regions by mean-shift algorithm. The region-based graph cut optimization algorithm is then adopted to minimize a novel energy function including the likelihood energy, prior energy and shape prior energy. The introduction of the shape prior energy term makes the foreground object contour coherent across all the images in the sequence. Finally, the pixel-based graph cut optimization within the narrow band of the coarse segmentation result can further refine the object boundary. The algorithm extracts foreground accurately and efficiently by combining the graph cut optimization technique with an intelligent user interface.

5. OVERVIEW OF OUR SYSTEM

We have developed a fully automatic multiple objects segmentation system from multi-view video sequence. The framework of the proposed system is summarized in Fig 9. It includes three major components: data acquisition, depth reconstruction, object segmentation and tracking. The system output, i.e. the object segmentation results, can be used for the object-based video rendering.

(a)control unit (b)camera unit

Fig.10: *Capturing system components and camera configuration.*

5.1 Data acquisition

5.1.1 Multi-view video system

The input video sequences are captured by a multi-view video system built in the Visual Signal Processing and Communications Laboratory of the Chinese University of Hong Kong. The system components and camera configuration are shown in Fig 10. The system includes five video cameras and a control unit. As shown in Fig 10(a), a control computer is connected to the camera units through the IEEE 1394 interfaces for data transmission. All the cameras are synchronized by a synchronizer. A hard disk array is used to store the captured video data. As shown in Fig 10(b), the five camera units are arranged along an arc spanning a small angle with approximately equal distance between neighbors, and face the center of the arc. Each camera set with a lens is mounted on the tripod. The model of the lens is *Prosilica GC650C* with 8mm fixed focal length. The model of the tripod is *Manfrotto 028/028B Triman Tripod*, whose pan-tilt can be adjusted in three directions.

During data capturing, even though the parameters for exposure and white balance have been carefully tuned, potential signal differences between images of different views are still inevitable, which introduce undesired effect for further steps. Thus in the post-processing step, color equalization is applied to the captured raw data to equalize the camera color responses across all the views.

5.1.2 Camera calibration

For the camera configuration shown in Fig 10(b), the five cameras (denoted by $C_0, C_1, \ldots, C_4$) are distributed roughly along an arc focusing on the scene. The camera coordinate system of C_2 is specified as the world coordinate system, thus its extrinsic parameters are $R = I$ and $t = 0$. The target of camera calibration is to calculate the fundamental matrices between interacting camera pairs, as well as the projection matrices of other cameras (C_0, C_1, C_3, C_4) w.r.t. the specified world coordinate system (C_2).

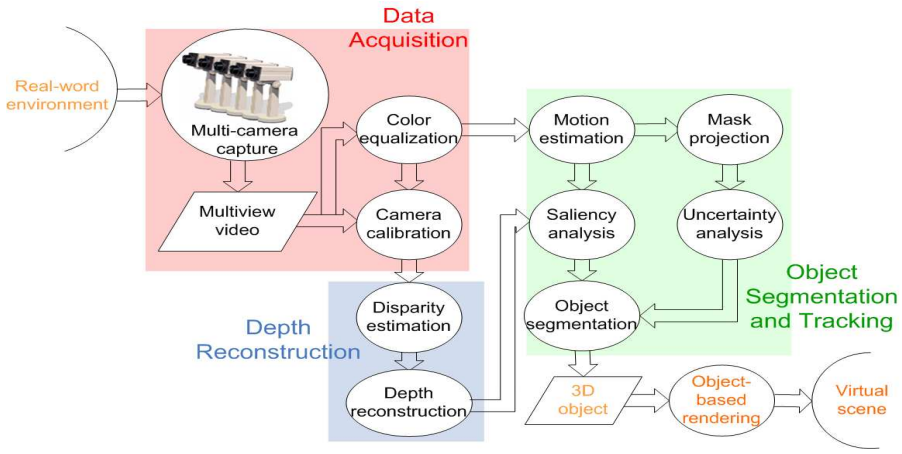


Fig.9: *Framework of the proposed system.*

To accomplish the multi-camera calibration, we first calibrate each individual camera separately using Zhang's method [23] as implemented in the Intel OpenCV library. Thus, the intrinsic matrices $A_i(i = 0, \ldots, 4)$ of each camera together with the plane position and orientation (R_i, t_i) relative to each camera can be obtained. However, to directly derive the rotation R and translation t between two cameras from the position and orientation of related plane is not robust and usually leads to noisy and erroneous results. Based on homography, we propose a nonlinear method to estimate accurate (R, t) between two cameras.

Suppose two cameras, C_1 and C_2 capture a planar pattern simultaneously as shown in Fig 7. A_1 and A_2 are their intrinsic matrices. The plane position and orientation expressed in the two different camera coordinate systems is given by (R_1, t_1) and (R_2, t_2). Let $m_1 = [u_1, v_1]^T$ and $m_2 = [u_2, v_2]^T$ denote the projections of the same 3D point on the planar pattern onto camera C_1 and C_2. $\overline{m_1} = [u_1, v_1, 1]^T$ and $\overline{m_2} = [u_2, v_2, 1]^T$ denote their homogeneous coordinates. The 3×3 homography matrix introduced by the plane can be estimated by a few point correspondences according to 6:

$$\overline{m_2} \simeq H\overline{m_1} \tag{6}$$

The homography is related to the intrinsic and extrinsic parameters as described in 7:

$$H \simeq A_1(R + t \cdot n^T) \tag{7}$$

where n is the plane normal. Accroding to 7, we have:

$$G = (R + t \cdot n^T) \tag{8}$$

Given p image pairs, p different $G_j(j = 1, 2, \ldots, p)$ can be obtained by the estimated homography and intrinsic matrices. Thus the rotation R and translation t can be computed by solving the nonlinear minimization problem formulated below:

$$\min_{R,t} \sum_j \| G_j - R - t \cdot n^T \|_F \; subject\ to\ R^T R = I \tag{9}$$

To impose the orthogonal constraint $R^T R = I$, we use the angle-axis representation of rotation with three DOF (degree of freedom). Thus the minimization problem can be solved using the Levenberg - Marquardt algorithm.

For global registration, the transform from any camera to other ones can be derived easily by chaining those estimated pair-wise transforms together across frames. Accurate and stable pair-wise calibration can improve the accuracy of chaining and our experiments show that there is no obvious error accumulation during the transform chaining using the proposed pair-wise extrinsic calibration.

Finally, the fundamental matrix between two cameras can be computed by 10 and the projection matrix of each camera can be recovered by 2 after registering the position and orientation of all the cameras to the world coordinate system.

$$F = A_2^{-T} R[t]_x A_1^{-1} \tag{10}$$

5.2 Depth reconstruction

In order to reconstruct the depth map, we propose a method to estimate the disparity fields using a regularization scheme with occlusion reasoning, where piecewise smoothness is enabled while discontinuities at object boundaries are preserved. Epipolar constraint recovered by camera calibration is used to convert the two dimensional matching problem into a one dimensional solution. Given the correspondences among multi-view images, depth map can be effectively reconstructed by solving the linear equations based on the perspective projection model of 2.

5.2.1 Occlusion reasoning

In the proposed method, the occlusion reasoning is derived directly from its definition that if a pixel cannot find a correspondence in the reference image, it is labeled as occluded. Let I_1 and I_2 represent the input image pair. Let O_{12} denotes the occlusion map of image I_1 w.r.t. the reference image I_2, and O_{21} denotes the occlusion map of I_2 w.r.t. I_1. Similar representation is used for the disparity maps, i.e., D_{12} and D_{21}. We then define three occlusion statuses for each pixel, namely, not occluded, definitely occluded and possibly occluded. The penalty values of the three statuses are set to OC_n, OC_d and OC_p, respectively.

Not occluded A pixel in image I_1 is not occluded $(O_{12}(p_1) = OC_n)$if there exists a corresponding pixel p_2 in image I_2, such that $p_2 + D_{21}(p_2) = p_1$. Not occluded pixels of O_{12} can be determined either during the disparity estimation process or by inverse warping of the disparity map D_{21}.

Definitely occluded Definitely occluded pixels are determined by the geometric constraint. According to the epipolar constraint, the corresponding pixel p_2 in image I_2 for pixel p_1 in image I_1 lies on the epipolar line $l_{12} = \mathbf{F}_{12}p_1$. However, line l_{12} may not have intersection with the image plane within the height and width limits of the x and y axes. In this case, the pixel p_1 in I_1 is labeled as definitely occluded.

Possibly occluded Possibly occluded pixels are derived from the definition of occlusion and are located during the disparity estimation or by view warping as a post-processing step. Given the disparity map D_{12}, each pixel p_1 in I_1 points to a certain pixel in I_2 through $D_{12}(p_1))$. As the uniqueness constraint is not enforced, multiple pixels in I_1 may point to the same pixel in I_2. Thus, there exist pixels in I_2 that are never being pointed to, and these pixels will be labeled as possibly occluded, i.e., $O_{21}(p_2) = OC_p$.

We make some important observations here. Firstly, as the definitely occluded pixels are determined by the calibration parameters of the multi-camera system, which remain unchanged once the images are captured, they are fixed once labeled. Therefore, during the regularization process, the definitely occluded pixels are determined at the first iteration and do not change in later iterations. On the other hand, the possibly occluded pixels are updated at each iteration.

Secondly, definite occlusion normally occurs at image boundaries, while possible occlusion often occurs at object boundaries. The occlusion map is also piecewise smooth, thus an edge-preserving smoothness term can be used to constrain the occlusion map. However, this observation does not apply to definite occlusion. When computing the smoothness term w.r.t. the neighbors, the definitely occluded neighboring pixels are excluded.

Thirdly, to avoid all the pixels being labeled occlusion, a penalty for occlusion should be set. In our algorithm, we assign a large value for OC_d and a small value for OC_p. OC_n is set to zero. Including the occlusion penalty in the overall energy function embodies the uniqueness constraint as a soft constraint. When calculating the disparity for pixel p_1 in I_1, as the energy function shall be minimized, the pixel p_2 in I_2 whose current status is possibly occluded will be preferred to the pixels who have already been matched by some other pixel in I_1, i.e., whose current status is not occluded. In other words, the proposed algorithm favors but not enforces uniqueness.

5.2.2 Two-view disparity estimation

The proposed algorithm is asymmetric in the sense that it calculates simultaneously the disparity of the target image D_{12} and the occlusion map of the reference image O_{21}. The proposed energy function is

$$E = E_d(D_{12}) + \lambda_d \cdot E_{ds}(D_{12}) + \gamma(E_o(O_{21}) + \lambda_o \cdot E_{os}(O_{21})) \tag{11}$$

where E_d is the disparity data term, and E_{ds} is the disparity smoothness term. E_o and E_{os} are the occlusion data term and the occlusion smoothness term, respectively. γ is a weighting factor to trade-off between disparity and occlusion terms. λ_d and λ_o denote the Lagrange parameters for disparity and occlusion regularization. Specifically, for a pixel p in image I_1 and a candidate matching pixel q in image I_2, the four terms in the energy function are defined as follows:

$$E_d(p, q) = \frac{1}{n} \cdot \sum_{(p_i \in N_p, q_i \in N_q)} dist(I_1(p_i), I_2(q_i)) \tag{12}$$

$$E_o(q) = -|O_{21}(q) - OC_n| \tag{14}$$

Note that in 11, the energy terms are defined on the disparity field D or the occlusion map O, while through 12 to 15, the energy terms are defined for a particular pixel p or a pair of pixels p and q. For convenience we do not differentiate between these two representations.

In the above energy functions, n is the total number of items counted in the summation. N_p defines the neighborhood of pixel p. The color distance between two pixels is defined as the mean absolute differences (MAD) for R, G, B components. Function f is used to suppress noise in the gradient image, where the gradients smaller than the threshold T are set to zero.

The disparity data term E_d measures the photo-consistency covering a small neighborhood of the target pixel p and its matching candidate q. As the denominator of the smoothness terms E_{ds} and E_{os} incorporates the image gradient information, the disparity and occlusion maps can be well smoothed

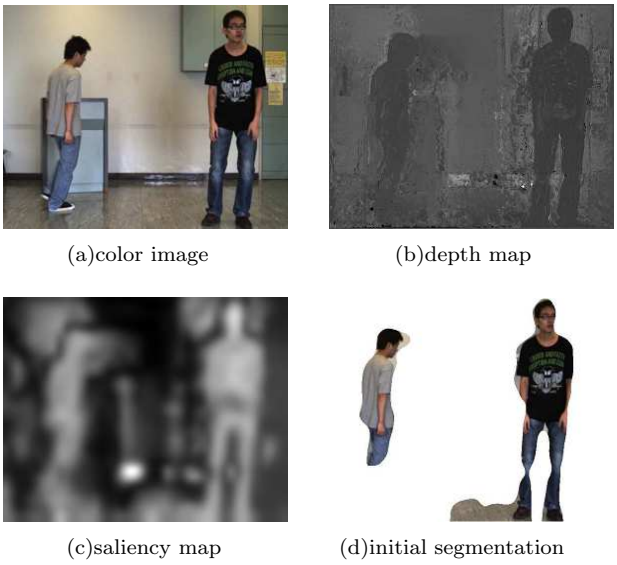
$$E_{ds}(p) = \frac{1}{n} \cdot \sum_{(p_i \in N_p, O_{12}(p_i) \neq OC_d)} \frac{|D_{12}(p_i) - D_{12}(p)|}{f(dist(I_1(p_i), I_1(p)), T) + 1} \tag{13}$$

$$E_{os}(q) = \frac{1}{n} \cdot \sum_{(q_i \in N_q, O_{21}(q_i) \neq OC_d)} \frac{|O_{21}(q_i) - OC_n)|}{f(dist(I_2(q_i), I_2(q)), T) + 1} \tag{15}$$

in textureless regions, while discontinuities are preserved at edges. Note that gradient of I_1 is used for disparity smoothness term while gradient of I_2 is used for occlusion smoothness term.

Note that during the 2-view disparity estimation procedure, when we calculate the disparity field D_{12} of image I_1, the geometric analysis is performed based on I_1 and thus the definitely occluded part of O_{12} can be obtained. However, by warping I_1 through D_{12}, it is O_{21} whose possibly occluded part can be obtained. The occlusion data term and smoothness term are based on O_{21} for image I_2. Thus, the proposed algorithm solves the disparity estimation problem in an asymmetric manner.

The novelty of the proposed algorithm lies in the terms E_o and E_{os}. The purpose of adding the occlusion data term is twofold. The first is to avoid labeling all the pixels as occluded. In addition, it enables the uniqueness characteristic as a soft constraint. The advantage of adding the occlusion smoothness term is to ensure piecewise smoothness within the occlusion map, avoiding holes and isolated occluded points.



(a)color image (b)depth map (c)saliency map (d)initial segmentation

Fig.11: *Saliency-based segmentation initialization.*

5.3 Automatic object segmentation from multi-view videos

5.3.1 Automatic segmentation initialization

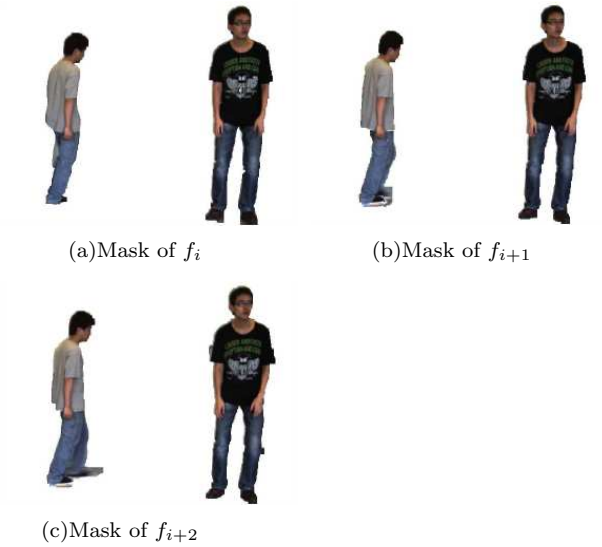
Automatically locating a semantic object in the visual environment is an effortless task for human observers but a challenging problem for machines. Generally, user interventions are required to extract interested objects from single image, and motion information is analyzed for object extraction in video sequences. However, the user intervention is inconvenient and the motion analysis can only extract moving objects. These drawbacks motivate the development of automatic and robust algorithm for object extraction. The visual attention concept provides an effective mechanism to simulate human's unconscious attention towards the locations of interested objects in a complicated scene. Itti and Koch [80] proposed a saliency-based visual attention model for scene analysis and attention location, which combines low-level image features, i.e., color, intensity and orientation. The value in the saliency map indicates the conspicuity, i.e., a large value attracts more attention. However, interested objects do not always comply with these straightforward low-level features, and hence the computed saliency map may not represent the visual attention well. For more efficient and robust performance, high-level features should be taken into consideration. In our system, the fully automatic segmentation initialization is realized by object-of-interest (OOI) extraction algorithm based on the visual attention model. We assume that interested objects should undergo continuous and smooth distribution on both depth and motion properties that can be obtained from multi-view videos. Thus in the proposed system, these two high level features are involved in the saliency map calculation for OOI extraction in the visual environment.

Based on the saliency map, OOIs can be located and employed to initialize the segmentation algorithm. Thresholding [81] is firstly applied on the saliency map. Then some noises are removed by morphological erosion and dilation operations. Subsequently, we use connected component labeling (CCL) approach to detect the location and measure the size of each component. After thresholding on the size of the components, those components with small magnitude are eliminated and the remaining large ones are considered to be the object patches. The number of the objects can be obtained simultaneously. Fig 11 contains the experimental results of segmentation initialization.

5.3.2 Graph cut based multiple object segmentation and tracking

Compared with the extensive interest and great accomplishment in binary label segmentation, multiple label segmentation is a less popular research topic because of its difficulty and complicity. In our system, we propose a depth-assisted multiple object segmentation algorithm from multi-view videos using graph cut. We follow the formulation of energy function defined in 5. Color, depth and motion cues are involved to calculate the data term. Given the results of segmentation initialization obtained in the saliency-based object extraction stage, we use Gaussian Mixture Model (GMM) to model the pixel color distribution. One GMM is built for each object as well as the background, where the GMM has 10 components for background and 5 components for foreground, respectively. For the depth and motion cues, histograms are used to model their distributions. For the choice of the smoothness term, since the depth and motion values may not be accurate at the object boundary, we only encode the color gradient to improve the overall performance. Then a local adaptive normalization in the pixel's second-order neighborhood system is imposed on the smoothness term, which guarantees that the weights of a pixel with high discontinuity distributions in its neighborhood can be suppressed while that with low discontinuity distributions enhanced. For the proposed energy function, multi-way cut with α-expansion is carried out for energy minimization.

Tracking is a key technique for good video segmentation performance. Available motion information between adjacent frames can be directly utilized in the tracking step. To segment the following frames, firstly, foreground objects of the previous frame are projected to the current frame to form the initial mask with the help of motion information. Then morphological operation is used to remove noises and small holes. After that, we locate the foreground objects by CCL and mark out the contour of each object. Searching along the object contours, we can define the definite foreground, definite background and uncertain areas. At last, the proposed multi-way cut algorithm is performed within the uncertain areas to segment out the objects for the current frame. In Fig 12, the segmentation results of three successive frames in our captured video sequences demonstrate the efficiency of our method.



(a)Mask of f_i (b)Mask of f_{i+1}

(c)Mask of f_{i+2}

Fig.12: *Segmentation result of three successive frames.*

6. CONCLUSION

In this paper, a tutorial of multi-view VBOS systems and the related techniques including data acquisition, depth reconstruction, object segmentation and tracking is given. We started with a brief review of typical multi-view video capturing systems developed in recent decades. We then introduced a number of representative camera calibration approaches and discussed their applicability for multi-view capturing systems. Next, a variety of stereo matching algorithms for depth reconstruction are reviewed according to different scene representation and optimization techniques. After that, classical and state-of-the-art graph-cut based object segmentation algorithms are presented. At last, we gave the framework of our multiple objects segmentation system from multi-view video sequence and described its related technical details to illustrate the practical implementation of a multi-view VBOS system.

7. ACKNOWLEDGEMENT

This work was supported in part by the Research Grants Council of Hong Kong SAR (Project CUHK415707).

References

[1] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, M. F. Cohen, "Interactive Video Cutout," *Proceedings of ACM SIGGRAPH 2005.,* vol. 24, pp. 585-594, 2005.

[2] Y. Li, J. Sun, H. Y. Shum, "Video Object Cut and Paste," *Proceedings of ACM SIGGRAPH 2005.,* volF. 23, pp. 595-600, 2005.

[3] J. S. Cardoso, J. C. S. Cardoso, L. Corte-Real, "Object-Based Spatial Segmentation of Video Guided by Depth and Motion Information," *IEEE Workshop on Motion and Video Computing.,* pp. 7-7, 2007.

[4] I. D. Reid, K. Connor, "Multiview Segmentation and Tracking of Dynamic Occluding Layers," *Proc. 16th British Machine Vision Conference.,* vol. 2, pp. 919-928, , 2005.

[5] P. J. Narayanan, P. Rander, T. Kanade, "Synchronous capture of image sequences from multi-

ple cameras," *tech. rep.,* The Robotics Institute, CMU, 1995.

[6] T. Kanade, H. Saito, S. Vedula, "The 3D room: digitizing time-varying 3D events by synchronized multiple video streams," *tech. rep.,* The Robotics Institute, CMU, 1998.

[7] C. Zhang, T. Chen, "Multi-View Imaging: Capturing and Rendering Interactive Environments," *Proc. Computer Vision for Interactive and Intelligent Environment.,* pp. 51-67, 2005.

[8] T. Naemura, J. Tago, H. Harashima, "Real-Time Video-Based Modeling and Rendering of 3D Scenes," *IEEE Computer Graphics and Applications.,* vol. 22, pp. 66-73, 2002.

[9] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. on Graphics.,* vol. 23, pp. 600-608, 2004.

[10] N. Inamoto, H. Saito, "Intermediate view generation of soccer scene from multiple videos," *IEEE Conf. on Computer Vision Pattern Recognition.,* vol. 2, pp. 713-716, 2002.

[11] W. Matusik, C. Buehler, R. Raskar, "Image-Based Visual Hulls," *Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH.,* pp.369-376, 2000.

[12] R. G. Yang, G. Welch, G. Bishop, "Real-Time Consensus-Based Scene Reconstruction using Commodity Graphics Hardware," *Proc. Pacific Conf. Computer Graphics and Applications.,* pp. 225-234, 2002.

[13] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, C. Zhang, "Multi-view Imaging and 3DTV," *Proc. IEEE Signal Processing Magazine.,* vol. 24, pp.10-21, 2007.

[14] A. Fusiello, "Uncalibrated euclidean reconstruction: a review," *Image and Vision Computing.,* vol.18, pp. 555-563, 2000.

[15] J. Salvi, X. Armangué, J. Batlle, "A comparative review of camera calibrating methods with accuracy evaluation," *Pattern Recognition.,* vol. 35, pp. 1617-1635, 2002.

[16] R. Y. Tsai, "A versatile camera calibration technique for high accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE J. Robotics and Automation.,* 1987.

[17] O. Faugeras, G. Toscani, "The calibration problem for stereo," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition.,* pp.15-20. 1986.

[18] J. Weng, P. Cohen, M. Herniou, "Camera calibration with distortion models and accuracy evaluation," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 14, no. 10, pp.965-980, 1992.

[19] J. Heikkilä, "Geometric camera calibration using circular control points," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 22, no. 10, pp.1066-1077, 2000.

[20] O. Faugeras, T. Luong, S. Maybank, "Camera self calibration: theory and experiments," *European Conf. on Computer Vision.,* pp.321-334, 1992.

[21] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision," *Cambridge University Press.,* 2002.

[22] S. Bougnoux, "From projective to euclidean space under any practical situation, a criticism of self-calibration.," *Proc. 6th Intl. Conf. on Computer Vision.,* pp. 790-796, 1998.

[23] Z. Zhang, "A flexible new technique for camera calibration," *IIEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 22, no.11, pp. 1330-1334, 2000.

[24] P. Sturm, S. Maybank, "On plane-based camera calibration: a general algorithm, singularities, applications," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition.,* pp. 432-437, 1999.

[25] B. Triggs, "Autocalibration from planar scenes," *European Conf. on Computer Vision.,* pp. 89-105, 1998.

[26] S. Prince, A. D. Cheok, F. Farbiz, T. Williamson, N. Johnson, M.Billinghurst, H. Kato, "3D Live: Real Time Captured Content for Mixed Reality," *International Symposium on Mixed and Augmented Reality.,* pp. 307-317, 2002.

[27] I. hrke, L. Ahrenberg, M. Magnor, "External Camera Calibration for Synchronized Multi-video Systems.," *Journal of WSCG.,* vol. 12, 2004.

[28] B. Caprile, V. Torre, "Using vanishing points for camera calibration," *International Journal of Computer Vision.,* vol. 4, no. 2, pp.127-140, 1990.

[29] P. Sturm, B. Triggs, "A Factorization Based Algorithm for Multi-Image Projective Structure and Motion," *European Conference on Computer Vision.,* pp. 709-720, 1996.

[30] T. Ueshiba, F. Tomita, "Plane-based Calibration Algorithm for Multi-camera Systems via Factorization of Homography Matrices," *International Conference on Computer Vision.,* vol. 2, pp. 966-973, 2003.

[31] T. Svoboda, D. Martinec, T. Pajdla, "A Convenient Multicamera Self-calibration for Virtual Environments," *PRESENCE: Teleoperators and Virtual Environments.,* vol. 14, no. 4, 2005.

[32] M. A. Penna, "Camera Calibration: A Quick and Easy Way to Determine the Scale Factor," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* , vol. 13, no. 12, pp. 1240-1245, 1991.

[33] D. Daucher, M. Dhome, J. Lapreste, "Camera

Calibration from Spheres Images," *Proc. European Conf. Computer Vision.,* pp. 449-454, 1994.

[34] H. Teramoto, G. Xu, "Camera Calibration by a Single Image of Balls: From Conics to the Absolute Conic," *Proc. Fifth Asian Conf. Computer Vision.,* pp. 499-506, 2002.

[35] H. Zhang, Y. Wong, G. Zhang, "Camera Calibration from Images of Spheres," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 29, no. 3, 2007

[36] X. Chen, J. Davis, P. Slusallek, "Wide Area Camera Calibration Using Virtual Calibration Objects," *IEEE Conf. on Computer Vision Pattern Recognition.,* vol. 2, pp. 520-527, 2000 .

[37] D. Scharstein, R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision.,* vol. 47, no. 1-3, pp. 7-42, 2002.

[38] T. Kanade, M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 16, no. 9, pp. 920-932, 1994.

[39] O. Veksler, "Stereo correspondence with compact windows via minimum ratio cycle," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 24, no. 12, pp. 1654-1660, 2002.

[40] L. Tang, C. Wu, Z. Chen, "Image dense matching based on region growth with adaptive window," *Pattern Recognition Letters.,* vol. 23, pp. 1169-1178, 2002.

[41] K. J. Yoon, I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 28, no. 4, pp. 650-656, 2006.

[42] Y. Zhang, C. Kambhamettu, "Stereo matching with segmentation based cooperation," *European Conf. on Computer Vision.,* pp. 556-571, 2002.

[43] L. Hong and G. Chen, "Segment-based stereo matching using graph cuts," *IEEE Conf. on Computer Vision Pattern Recognition.,* pp.74-81, 2004.

[44] L. Zitnick, S. B. Kang, "Stereo for image-based rendering using image over-segmentation," *International Journal of Computer Vision.,* 2007.

[45] Y. Taguchi, B. Wilburn, L. Zitnick, "Stereo reconstruction with mixed pixels using adaptive over-segmentation," *IEEE Conf. on Computer Vision Pattern Recognition.,* 2008.

[46] M. Gong, R. Yang, "Image-gradient-guided real-time stereo on graphics hardware.," *Proc. IEEE 3DIM.,* pp. 548-555, 2005.

[47] S. Yoon, D. Min, K. Sohn, "Fast dense stereo matching using adaptive window in hierarchical framework," *Proc. Int. Symposium on Visual Computing.,* pp. 316-325, 2006.

[48] V. Kolmogorov, R. Zabih, "Computing visual correspondence with occlusion using graph cuts," *Proc. of International Conference on Computer Vision .,* pp. 508-515, 2001.

[49] V. Kolmogorov, R. Zabih, "Multi-camera scene reconstruction via graph cuts," *Proc. of European Conference on Computer Vision.,* pp. 82-96, 2002.

[50] Y. Wei, L. Quan, "Asymmetrical occlusion handling using graph cut for multi-view stereo," *Proc. Computer Vision and Pattern Recognition.,* pp. 902-909, 2005.

[51] J. Sun, N. N. Zheng, H. Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence.,* vol. 25, no. 7, pp. 787-800, 2003.

[52] A. Klaus, M. Sormann, K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," *International Conference on Pattern Recognition.,* 2006.

[53] S. Larsen, P. Mordohai, M. Pollefeys, H. Fuchs, "Temporally consistent reconstruction from multiple video streams using enhanced belief propagation," *IEEE International Conference on Computer Vision.,* 2007.

[54] M. F. Tappen, W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," *IEEE International Conference on Computer Vision.,* 2003.

[55] P. F. Felzenszwalb, D. P. Huttenlocher, "Efficient Belief Propagation for Early Vision," *International Journal of Computer Vision.,* vol. 70, no. 1, pp. 41-54, 2006

[56] Y. Ohta, T. Kanade, "Stereo by intra- and inter scanline search using dynamic programming," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 7, no. 2, pp. 139-154, 1985.

[57] P. N. Belhumeur, "A Bayesian approach to binocular stereopsis," *International Journal of Computer Vision.,* vol. 19, no. 3, pp. 237-260. 1996.

[58] P. N. Belhumeur, D. Mumford, "A Bayesian treatment of the stereo correspondence problem using half-occuluded regions," *IEEE Conf. on Computer Vision Pattern Recognition.,* pp. 506-512, 1992.

[59] A. F. Bobick, S. S. Intille, "Large occlusion stereo," *International Journal of Computer Vision.,* vol. 33, no. 3, pp. 181-200, 1999.

[60] C. Kim, K. M. Lee, B. T. Choi, S. U. Lee, "A dense stereo matching using two-pass dynamic programming with generalized ground control points," *Proc. IEEE Conference on Computer Vision and Pattern Recognition.,* 2005.

[61] O. Veksler, "Stereo Correspondence by Dynamic Programming on a Tree," *Proc. IEEE Conference on Computer Vision and Pattern Recognition.,* vol. 2, pp. 384-390, 2005.

[62] "vision.middlebury.edu/stereo/".

[63] D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 8, no. 4, pp. 413-424, 1986.

[64] S. Geman, D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 6, no. 6, pp. 721-741, 1984.

[65] S. T. Barnard, "Stochastic stereo matching over scale," *International Journal of Computer Vision.,* vol. 3, no. 1, pp.17-32, 1989.

[66] D. M. Greig, B. T. Porteous and A. H. Seheult , "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society Series B.,* vol. 51, pp.271-279, 1989.

[67] Y. Boykov, M. P. Jolly, "Interactive Graph cuts for optimal boundary and region segmentation of objects in N-D images," *Proc. IEEE Int. Conf. Computer Vision.,* pp: 105-112, 2001 .

[68] L. Ford, D. Fulkerson, "Flows in network.," *Princeton University Press.,* 1962.

[69] Y. Boykov, O. Veksler, R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 23, no. 11, pp: 1222- 1239, 2001.

[70] Y. Li, J. Sun, C. K. Tang, H. Y. Shum, "Lazy Snapping," *ACM Tranc. Graph.,* vol. 23, pp: 303-308, 2004.

[71] L. Vincent, P. Soille, "Watersheds in digital space: an efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 13, pp: 583-598, 1991.

[72] C. Rother, V. Kolmogorov, A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Tranc. Graph.,* vol. 23, pp: 309-314, 2004.

[73] S. Wang, J. M. Siskind, "Image Segmentation with Ratio Cut," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 25, pp: 675-690, 2003.

[74] J. B, Shi, J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 22, pp: 888-905, 2000.

[75] Y. Mu, H. Zhang, H. L. Wang, W. Zuo, "Automatic video object segmentation using graph cut," *Proc. IEEE Int. Conf. Image Processing.,* vol. 3, pp: 377-380, 2007.

[76] S. F. Chen, L. L. Cao, J. Z. Liu, X. O. Tang, "Iterative MAP and ML Estimations for Image Segmentation.," *IEEE Conf. Computer Vision and Pattern Recognition.,* pp: 1-6, 2007.

[77] Y. P. Tasi, C. H. Ko, Y. P.Hung, Z. C. Shih, "Background Removal of Multiview Images by Learning Shape Priors," *IEEE Trans. Image Processing.,* vol. 16, pp: 2607-2616, 2007.

[78] B. Goldlücke, M. A. Magnor, "Joint 3D-Reconstruction and Background Removal Separation in Multiple Views using Graph Cuts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition.,* vol. 1, pp: 683-688, 2003.

[79] M. Sormann, C. Zach, K. Karner, "Graph Cut Based Multiple View Segmentation for 3D Reconstruction," *Proc. Int. Symposium on 3D Data Processing, Visualization, and Transmission.,* pp: 1085-1092, 2006.

[80] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence.,* vol. 20, pp: 1254-1259, 1998.

[81] C. Strouthopoulos, N. Papamarkos, "Multithresholding of mixed type documents," *Engineering Application of Artificial Intelligence.,* vol. 13, no. 3, pp: 323-343, 2000.

ChunHui Cui received his B.E. and M.E. degree from Huazhong University of Science and Technology, China, in 2004 and 2006, respectively. He is currently a Ph.D. Candidate in the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong. His current research interests are Image Processing and Computer Vision.

Qian Zhang was born in Xi'an, China in 1982. She received the M.Sc degree in Computer Science and Technology from XiDian University, Xi'an, China , in 2007. Now, She is a Ph.D. candidate in the Department of Electronic Engineering, the Chinese University of Hong Kong. Her research interests include Multimedia Communication and Image Segmentation.

KingNgi Ngan received his B.Sc. (Hons) and Ph.D. degrees, both in Electrical Engineering from Loughborough University, U.K., in 1978 and 1982, respectively. He joined the Department of Electronic Engineering, the Chinese University of Hong Kong as Chair Professor in 2003. Recently, he has been appointed as IEEE Distinguished Lecturer of the Circuits and Systems Society.