

## Video image-based intelligent architecture for human motion capture

Lei Jun<sup>1</sup>, Dieter Hogrefe<sup>1</sup>, Tan Jianrong<sup>2</sup>  
1 Telematics Group, University of Goettingen  
37083 Goettingen, Germany  
Email: [lei.hogrefe@cs.uni-goettingen.de](mailto:lei.hogrefe@cs.uni-goettingen.de)  
<http://www.tmg.informatik.uni-goettingen.de>  
2 Graphics Institute Zhejiang University  
310009 China  
Email: [egi@zju.edu.cn](mailto:egi@zju.edu.cn)

**Abstract:** The study of human motion is a fascinating subject in computer vision and has been developed quite a few years. Traditional methods to realize human motion capture mainly use equipments of machinery, electromagnetic, acoustics, optics, graphics and so on. Whereas these are some defects existed in these methods e.g. expensive equipments, the place limited by sensors and localization of sport range, and it has no consistent standards because of different purposes and requirements. Most importantly, they always need a person to execute the long and sterile work for surveillance. Sometimes only by manual operation cannot accord with desires that data should be collected without redundant or unhelpful information. In view of video characters which bring about changes and new opportunities in the human motion capturing technology, this paper presents an actual architecture to automatically capture human motion by comparing video images and extracting surveillant area. Then we provide an effective method for storage. Furthermore, the functionalities of architecture are realized in GOLF sport education example.

**Key words:** *motion capture, video image, surveillant area, storage, video stream*

### 1. Introduction

Human motion is one of the most complex and compelling problems in computer vision today. Given the ability to recognize human motions and activities, computer vision systems prove enormously beneficial in a broad range of application areas including athletics and biomedicine and civil and defense security systems. Before such systems become reality, however, many challenging problems must be solved. In communication times, network and high-tech products bring about challenges and new opportunities in human motion capturing technology.

#### 1.1 Main techniques used in human motion capturing

The development of multimedia technologies induces data types from plane to 3D; from pure-text to the multimedia files including video and audio streams; from single frame image to continuous dynamic information. Video can be called image video because every frame of image in its sport sequence is composed of real-time collected natural visions or the digital objects converted from active objects [1]. Moreover the name of "Video Capture" is not accurate because it includes not only the capture but also a series of managements applied to the collected video. Only when the compression of video frequency, the conversion of color system, the rejection of noises and some necessary steps are taken into account, the video stream can be formed as a standard video file. Commonly a video file is formed by the visible image and the audible audio which is accessorial to be combined with image in postsynchronization [2].

Several years before human motion information has been noted as the number or simple sign, then as low-speed object, the high-speed object and split second phenomena with the help of computer. These data will be usually collected in a film by high-speed camera or video recorder. Then the images with human motion can be observed, compared and analyzed [3]. But they are not fit for the requirements of modern society that should be more accurate, more convenient to obtain the adjustable data. With the employment of digital production, the human motion information converts from simple numbers, signs to motion images, video which is easier to be accepted by people than images and applied to lots of areas because of its animation, veracity and convenience [4]. So human motion capturing depends more and more on video. We choose video image as our architecture basis just for the reason too.

Forming a digital video needs 3 parts of

cooperation based on hardware platform [5]: the equipments of putting out analog signals e.g. video recorder, TV or computer; the devices of capturing, converting or coding the analog signal e.g. special video capture cards; equipments of receiving and recording the digital data after coding e.g. multimedia personal computer (MPC). In process the most important device is the video capture card which provides the interfaces connecting with the equipment of analog video and computer, even the ability of converting the analog signal into digital data.

As the popularity of personal computer (PC) and nonlinearity editing technology, DV capture and the technology of digital video production become easier to available. Many researchers use Media Tools to realize the DV capture included the following functions [5]: 1) scanning automatically from DV tap and making into AVI or WAV files; 2) scanning automatically the whole tap and creating a group of capturing list; 3) collecting from the list and adjusting the points of putting in and out; 4) collecting single frame of image and storing into AVI file; 5) putting out the materials of capturing.

1.2 Basic structure and contents of the paper Follow a model-based approach in this paper we study technology of how to automatically capture human motion. Then we present a general architecture for capturing human motion and model one representative system based on video images. Due to the specific characters of GOLF sport education, the models are developed to provide an efficient and automatic technology of capturing the human motion on the basis of contrasting video images and extracting logical surveillant area. When it comes to the storage process, we design a new model in order to suit the requisition of capturing with 1 or 2 seconds data before storing.

After shortly reviewing the related work in Section 2, we present a general architecture of capturing motion human in Section 3, followed by an actual method of capturing effective data based on video images in Section 4 and an effective method of storing the captured information in Section 5. We summarize our modeling experiences and outline future work in Section 6.

## 2. Related Work

Multimedia technology developing day by day causes people pay close attention to the methods of capturing and editing the video. The video capturing from analog signals to digital signals, the quality of signals lies not only on the platform of hardware and software but also on the capabilities of video devices. For that video capturing is not the single step of capturing, the captured data have to be handled with a serious of extra works, e.g. compression of video, conversion of color, rejection of noise. Then data can be formed into normal digital video files and conserved for a long time.

Similar with acquiring the digital images, the sources formed into digital video file mainly have 3 ways:

I. Using the dynamic data from computer, e.g. converting FLC or GIF into the one type of videos--AVI

II. Static images or graphical files building up the sequences of video file

III. Adopting the method of video capture card; converting the analog signals into digital signals then saving the file with digital signals into the disk

As digital camera, DV and some high-tech productions force the realization of laconic, multilateral and personal methods for video capture. Video capture is also relative to different television standards.

At present, the television standards of each country may be different which is related to how to play video files with different decoders. There are three popular television standards types in the world: NTSC, PAL and SECAM, concerning about difference of frame rate, decomposed frequency, signal bandwidth and carrier frequency, the changed relation of color space [6].

NTSC is established by American standard committee of TV in 1952. PAL is another colorcast criterion, established by West Germany in 1962, which overcome the shortcoming of color distortion produced by phasic sensitivity of NTSC. SECAM is adopted in 1967 and has 625 lines and 25 frames per second. China and Europe adopted the type of PAL, while America and Japan adopted the type of NTSC. When we store the dynamic data, the television standards should be considered or the captured file can not be replayed normally.

Then how to capture human motion in general is presented as following. Depended on the human's long-time supervision, it is a type of numerous and baldness work to capture dynamic data with traditional methods. If we take this method, we should firstly check if the status of the equipments of video capturing is OK; if the output and input interfaces of video are in reason. Then we should frequently push down the buttons of "record" and "stop" when it is needed. When we are in the state of recording, it should be stopped in advance if we don't want to keep the previous part of data and video file would be wholly deleted, even those we needed. In process of long-time recording, many free time materials between the meaningful data would be kept down without the frequent exam and modification. These worthless data not only take up the capacity but also hinder the subsequent analysis unless done with extra refinement.

## 3. A general architecture of capturing human motion

Firstly we take advantage of video capture devices and adopt the general technology of capturing digital data, based on the technique of

DirectShow (which is the part of DirectX software package) [7]. The tool of “Graphedit” is often used to examine the connection during the capturing because it’s difficult to inspect the interfaces between hardware and software when we are in the state of software platform. We realize the digital video capturing by the following steps of connection in Laptop with IEEE1394 video capture card and general DV recorder:

I. Using the construct function to initialize the related information and setting up the basis of filter graph; initializing the class members including the interfaces with hardware, video frequency and video windows

II. Establishing the foundation of “graph” function for that every graph is the basis of each application which can be used to add other filters with different requirements

Only when the interfaces are all normal, the upstream filter and the downstream filter can be connected. Then we can utilize some functions of DirectShow software package to deal with the video stream.

III. Starting to connect the interfaces of devices and complete the integrity filter graph of capturing with the help of “Graphedit”.

Interfaces for DV recorder or video cam-recorder (VCR) are put into the graph with a counter. Then the container of Smart Tee (which is a part of DirectShow and provides the filter for connection with hardware and software) connects the hardware equipments into the “graph” whose interfaces join with the container of Splitter (which is a part of DirectShow and split the audio and video signal in video). The interfaces which follow Splitter are separately served for video and audio by way of mixing the separated audio and video by Avi Mux (which is a part of DirectShow). In the end the whole flow is to join with the interface of File Writer (which is used to create a video file). If we want to use the file successfully, some video decoders are needed, e.g. Dvcodec, Mpeg4Codec, and media player, e.g. RealPlayer. The overall flow is shown in Fig.1.

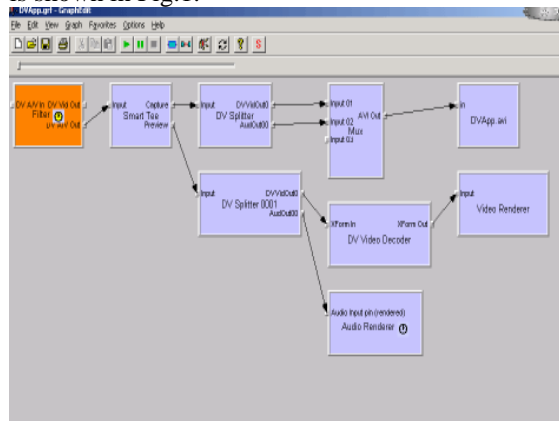


Fig.1. A general architecture of capturing

Now the connection is completed and the interfaces between the hardware and others can be examined. If the connections between all containers

are in gear and the input and output of streams are conformed to standards, we can start to capture video including human motion [8]. In Fig.1 we can see the two sessions for video stream: Preview and Capture. When we only use the function of preview, the video stream will not be recorded. Using DirectShow software package to capture the digital data can avoid the trouble caused by video for windows (VFW) software package because the signal collected by VFW needs to be converted, and improve the efficiency and quality [9].

The human motion can be noted in the sort of video types, e.g. MPEG, AVI or RM. By adopting the general architecture of capture and the help of human’s long-time supervision, we can obtain data including human motion. Otherwise man has to frequently press the buttons of “record” and “stop” if we don’t implement other methods. In the following contents, we will add a more rational method to the generic capturing system.

## 4. A system for automatically capturing human motion based on video images

### 4.1. The origin of video images comparison

The technology of automatic capturing adopts the method of video images comparison which mainly comes from the following ideas [10]:

I. When an object passes through a certain area, the gray level of pixels in this area will have a wave and then become stable. The value of changed gray level of pixels in this area will be almost the same as previous one.

II. When an object passes through a certain area, the gray level of pixels of this area will have a wave and then become stable too. The value of changed gray level of pixels in this area will be different.

III. If the lightness in this area changes the gray level of pixels will have smooth variant tendency.

Just for video being a combination of frames, we can construct a corresponding relationship by comparing the latter frame with the previous one. By the comparison, we can also confirm the sport model and orbit of the object.

During applying the method two values is needed to count: the variant difference between gray levels of pixels and the stable value of gray levels of pixels after change [9]. The previous value is put to examine if there is an object passed through, while the latter is used to judge if the object stops in this area. Then we will introduce the effective technology for extracting surveillant area based on images comparison.

### 4.2 Video image-based extracting surveillant area

In order to solve the problem of human’s long-time supervision and storing invalid information during the supervision; the automatic supervision of the effective area needs to be added into the general architecture of capturing. We assume the space of performing human motion capturing to be a full surveillant area. In this model, we mark up an area as the full surveillant area

which can hold the integrated human motion space in the vision. To get the aim of accuracy it is important to define a logical surveillant area, in which contains the subject for our surveillance to decide if the human motion should be captured or not.

To be convenient we set an example GOLF sport as our basis. During the process of GOLF sport, we only care about how the ball is stroked and the action of human. So we make the logical surveillant area smaller into  $4 \times 4$ , which can just hold the image of the ball in the vision (See Fig 2).

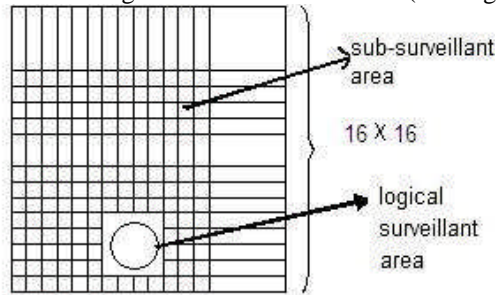


Fig.2. Extracting logical surveillant area

The smaller blank space is the logical surveillant area for GOLF ball because we only have to inspect the smaller space holding the image of GOLF ball when we use DV camera to capture effective human motion. Once the ball is stroked, the output equipment will receive a signal to start recording via the comparison of gray level of pixels in this area. Otherwise the video streams will not be written down either the ball isn't stroked or the gray level of pixels of this area doesn't change. The logical surveillant area isn't fixed and can be scalable according to the different requirements. The details how to confirm the logical surveillant area by requirements are shown in following section.

#### 4.3 The method of deciding the logical surveillant area

The surveillant area is divided into  $n \times n$  (in this model, choose  $n=16$ ) and every smaller one is a sub-surveillant area (See in Fig 2), which has the diverse coordinate number. For example, the second line and third area is marked as (2, 3) which is used in video image comparison. Certainly, we can choose the separated area being one of 256 portions or the combination of 2~3 portions as a surveillant area. These areas are discriminated by a hierarchy of sensitive degree. If we focus mainly on a certain area which can be marked as 0 (the highest of sensitive degree), others can be respectively marked by different degrees according to different attention levels. If an object appears in this area of degree 0, the capturing equipments start recording at once. Whereas the object appears in other areas of lower degree, it needs the second judgment with the second gray level consideration to decide if the video stream should be saved.

It is necessary to select the similar algorithm for extracting the major data because the rectangular divided area we want to use might doesn't suit the requirements. We can adopt the algorithm of the peripheral or ring encirclement in the area of the inspected object. Taking the encirclement as example, we mark the top, bottom, left and right as a, b, c and d by which we can get the corresponding portions. If these 4 points are all in one portion, the logical surveillant area is the current portion; if these portions border with each other, the logical surveillant area is the total; if these portions are separated, the logical surveillant area is determined by the following steps according to the coordinates of these 4 points: firstly the centre of inspected object needs to be decided by similar algorithm; then central area is determined by the centre which is almost the centre of the divided portion (an intersection point of rectangular diagonals) nearest to the centre of object; lastly the centre of the central area is taken as the center of a circle, the length of objects diagonal is taken as diameter. Every portion intersectant with this circle are included to the effective parts of the encirclement. The total portions are concerned as the encirclement of logical surveillant area.

We have expounded the model of how to touch off the equipment and start recording by taking advantage of the effective method, with which we can arrive at the aim of automatically supervision and capturing the useful data noting with human motion. In the following section, we concern about the effective method for storing the collected data.

### 5. An effective method of storing captured information

In the other side the storage of captured information is very important and directly determines the quality of human motion information mainly decided by different requirements. Real-time way needs real-time storage method and analysis after capturing can adopt mutual delayed-storage. In this architecture we construct a mutual delayed storage named "Tennis" storage because the working manners of two storage areas are very like playing tennis balls. (See Fig 3)

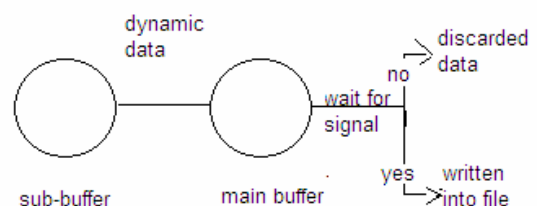


Fig.3. The process of "Tennis" storage

As video is the special type data, the real-time data can be written in the disk only when the

transmission rate of disk is larger than or same as the rate of captured data. Otherwise, the written file will lose some needed data. So Random Access Memory (RAM) is the primary choice to store the real-time data temporarily. If they are confirmed to be noted in disk, these data will be written into a file so that the transaction of real-time data can be avoided.

RAM is a medium to store temperate data which has smaller capacity than hard disk but can not store the data forever [11]. Some applications will take up a part of virtual memory or disk if the capacity is not enough. When we refer to the capturing human motion into a video file, the video memory is the main medium like RAM [12]. The video memory of computer for capturing is normally greater than 64M.

It's necessary to get the data of 1 second before stroking the ball if the storage for dynamic data of GOLF education is taken as an example. These data contain the movements of swing before stroking the ball and are most important to GOLF education. For the sake of 3~4 seconds of dynamic information being stored in an effective file, the data buffers should be created to store these temperate data. We establish 2 buffers in the video memory: the main buffer and the sub-buffer (See Fig 3) whose sizes are allocated by the concrete requirements of storage.

In this model, we allocate 2M to each buffer due to the large capacity and high-quality data of digital video. Every buffer can store almost 1s dynamic data. Then these temperate data can be deposited in buffers. When the main buffer is full of data, the sub-buffer will continue to be stored with data. When the sub-buffer is filled with data too, these data in the main buffer will be deleted and data in the sub-buffer will be moved to the main one. Then the whole process still carries on as previously. But when the signal for capturing the dynamic data stored into file arrives, the data in main buffer start to be written into file by File Writer (which is referred in Section 3) and wait for the signal for stopping. Then the information former 1 or 2 seconds can be written in the file together with the latter data. By this method it can not only improve the speed of capturing, but also save the time of CPU.

Certainly, we should take notice of saving the RAM to complete the procedure. These following methods can be adopted:

#### I. Setting free the most RAM as possible

Before video capturing, other applications and processes should be closed to the greatest extent, e.g. the program of screensaver. In this way, the video capture program can occupy with most RAM and reduce the loss rate.

#### II. Video capturing and compression in the same time

If the function of compression can be provided in the program, it should be used during

capturing dynamic data. By this way the pressure of RAM and disk will be reduced.

#### III. Using the optimize software to manage RAM

Before capturing, we can use the optimize software to manage RAM which can accelerate the speed of reading from and writing into RAM in order to suit the demand of capturing.

### 6. Summary our modeling experiences and outline future work

Take advantage of the effective video capturing technology, a proper model is constructed and apply into the GOLF education which is shown in Fig 4. It is certificated as a logical architecture for capturing the effective human motion.

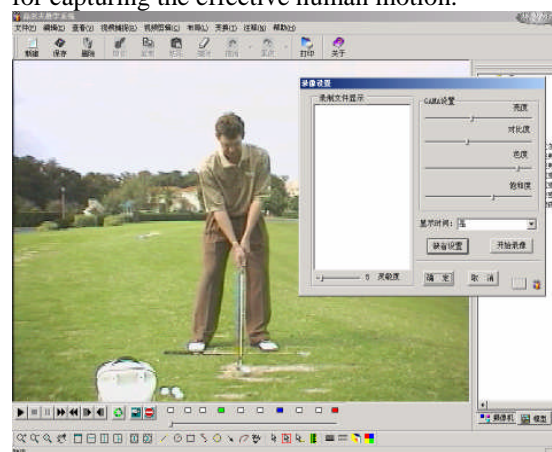


Fig.4. The architecture applied in GOLF education

This paper mainly concerns about the technology for capturing effective human motion and the corresponding work about extracting surveillant area and an effective storage method. We have introduced the video data, contemporal dynamic data and the basis of human motion capturing. The key of our model is extracting the logical surveillant area and set the sensitive degree including the comparison of gray level. Moreover, we adopt "tennis" storage model to improve the efficiency of capturing and save the capacity of disk. By this way, we arrive at the aim of automatically capturing useful human motion.

Now we can realize the real-time capturing and effective storage of human motion. Then we have applied the technology into the GOLF education and got satisfied results. Later we will take further look at the more effective and economic method. We believe that the technology of capturing will have further development with the renovation of digital production.

### Acknowledgment

We thank Dingli Communication Company in Zhuhai City and Computer Graphics Center of Zhejiang University in China for the technical support.

### Reference

- [1] William S. Managing Electronic Records (second edition). Prairie Village, Kansas. ARMA, 1998.
- [2] Tekalp A.M. Management of digital video. Beijing: Publishing House of Electronics Industry, 20~30, 1998.
- [3] Li HF. Digital medium---technology, application and design. Publishing House of Tsinghua. Apr. 2003.
- [4] Zhang WM, Wu LD. Multimedia information system. Beijing: Publishing House of Electronics Industry, 2002.
- [5] Zhong WZ, Li SQ. The technology of multimedia., Beijing: Publishing House of Tsinghua, 1993.
- [6] Ruan QQ. How to manage the digital image. Beijing: Publishing House of Electronics Industry. 2001.
- [7] Bargaen B, Donnelly P. Inside DirectX. Microsoft Press, April 1998.
- [8] Ren GJ, Zhang YL. Approach to real-time video image capture. Computer Engineering, 28: 268-270. 2002.
- [9] Video Capture on Windows. Geraint Davies Consulting Ltd. 2000.
- [10] Jiang HT, Helal A. Scene change detection techniques for video databases systems. Multimedia Systems, 6: 186~195. 1998.
- [11] Chang E, Zakhor A. Disk-Based Storage for Scalable Video. IEEE Transactions on circuits and systems for video technology; 7: 758~770. 1997.
- [12] Kanopoulos N, Hallenbeck J. J. A First-In, First-Out Memory for Signal Processing Applications. IEEE transaction on circuits and system for video technology, 33: 86-99.1986.



**Jun Lei** received the B.S degree in Mathematic Institute from Teacher-training college of Hangzhou, China, in 2001. She received M.S degree in Computer Graphics Institute from Zhejiang University, China in 2004 and is currently working towards the Ph.D. degree with Institute for Informatics, Goettingen University, Germany.

She entered Computer Graphics co. in Hangzhou, working as

TECHNICAL MANAGER from Mar.2002 to Oct.2002 and Attended UTStarcom co. Ltd P&T department, working as

TECHNICAL RESEARCHER from Nov.2002 to Feb.2003. She worked as TECHNICAL RESEARCHER in "Ding Li Mobile Communicate Company" from Oct.2003 to Dec.2003 in Zhuhai, China. Her research interests include design and analysis multimedia systems, digital image processing and now focus on multimedia distribution over wireless and wired broadband Internet access networks and Multiple Protocol Label Switching (MPLS) backbones and survivability, protection algorithm of optical WDM mesh networks.



**Dieter Hogrefe** graduated from Philips Exeter Academy, USA in 1976. He studied Computer Science and Mathematics at the Germany with a Diploma and University of Hannover, Germany with a Diploma degree and received his Ph.D in Computer Science in 1985.

His research activities are directed towards Computer Networks and Software Engineering. He has published numerous papers and two books on analysis, simulation and testing of formally specified communication systems. From 1983 to 1986 he was with the Siemens research centre in Munich and worked in the area of analysis of telecommunication systems. He was responsible for the protocol simulation and analysis of the CCS No. 7. He had professorships at the University of Hamburg, Bern, Luebeck and since 2002 he is FULL PROFESSOR (C4) for Telematics at the University of Goettingen.

Prof. Hogrefe represents the IITB (Frauhofer Institute for Information and Data Processing) in the European Telecommunication Standards Institute, ETSI, where is chairman of the Technical Committee Methods for Testing and Specification.



**Jiangrong Tan** received his Ph.D. degree in Mathematics Institute from Zhejiang University, China in 1992. Now he is the PROFESSOR and doctoral supervisor of Zhejiang University, China. His research interests include CAD, CIMS, virtual reality and scientific visualization, etc. He has published several papers and three books on CAD and CIMS. He is currently also the VICE

DIRECTOR of "state key lab of CAD&CG, Zhejiang University".

Prof. Tan received "the State Foundation for Excellent Young Scholars" in 1994 and "National Foundation on Graphics Science for Excellent Young Scholars" in 1995. He is also confirmed as "The leader of key study of Zhejiang Province" in 1997 and received "Top Grade Award in Excellent Teacher of BAOGANG" in 2003. He has also received the "State second award of science and technology" in 2004.