Chapter XX
Rationale, Design and Implementation of a Computer Vision–Based Interactive E–Learning System

Richard Y. D. Xu
Charles Sturt University, NSW, Australia

Jesse S. Jin
University of Newcastle, NSW, Australia

ABSTRACT

This article presents a schematic application of computer vision technologies to e-learning that is synchronous, peer-to-peer-based, and supports an instructor’s interaction with non-computer teaching equipments. The article first discusses the importance of these focused e-learning areas, where the properties include accurate bidirectional interaction and low cost hardware; system portability and versatile vision technology are emphasized. In the subsequent sections, we present some results aiming to achieve these goals. In particular, we highlight the most recent advancements in the interactive PTZ camera control from both the instructor and remote student. We also illustrated how these results have successfully addressed the challenges.

INTRODUCTION

In recent times, there has been an increase in the research activities aiming to apply computer vision (CV) technologies to various automated e-learning multimedia systems. In this article, we will discuss our progressive work in Intelligent Video Detection Agent (IVDA), which is a scheme of hardware design and computer vision software algorithms to assist e-learning systems that are synchronous, peer-to-peer based and an instructor uses non-computer based teaching equipment.
Before we present our prototype system, we need to explain the importance of these three properties, and illustrate how it differs from the other existing CV e-learning systems’ focuses.

### Synchronous E-Learning

In synchronous e-learning, the video and teaching multimedia are exchanged between the students and instructors in real-time, creating a highly interactive teaching and learning environment. Synchronous e-learning provides the students with instantaneous feedbacks and provides the instructor with a platform from which to monitor and adapt to student’s activities (Soreanu & Saucan, 2003).

Computer vision technologies are used in various synchronous e-learning systems to replace many manual, laborious and time-consuming tasks, which made real-time automated camera control and instantaneous multimedia authoring possible.

### Peer-to-Peer E-Learning

E-learning application can be classified according to its participant numbers, into peer-to-peer (P2P) and institutional ones. In P2P e-learning, there is usually one instructor and one student. Both participants can be of a variety of types, including home-based computer users. The learning can be both formal and informal.

Our work is focusing on P2P e-learning, since it allows for various types of participants, more importantly, the average computer users to take part in either or both the student and instructor role. It is an inexpensive learning platform and many studies can be identified to argue its advantages, such as literature in Jokela (2003) and Fletcher (2004).

However, the widespread of P2P e-learning participants also means that there will be additional considerations that need to be taken into account when designing computer vision software to support it. These considerations include:

1. The CV software must be able to intelligently assist the instructor when he works alone.
2. The CV software must adapt to standard equipment of an average computer user.

### Non-Computer Based Interaction

In the current synchronous e-learning applications, the instructor and student(s) usually communicated via standard computer interfaces, using chat window, computer whiteboards, sharable text and drawing pads, for example (Deshpande & Hwang, 2001).

However, many researchers have argued that e-learning cannot replace traditional learning altogether. In a similar argument, Liu & Kender (2004) mentioned that, based on a recent survey carried out at University of South Carolina, students consider traditional blackboard presentation as “essential” and “indispensable.”

Therefore, allowing the instructor, at least, to use traditional teaching equipments and video-capturing and streaming such information to student in real-time becomes a compromise solution between modern technology and traditional pedagogy.

However, most of the operator-less instructional video-capturing is achieved through the use of static camera(s) without computer vision processing. While this method is sufficient for some applications, the crude streaming suffers from an uninteresting presentation of teaching materials and is dependent on the instructor’s verbal instruction and remote student’s visual perception to determine what the student should pay attention to. In addition, the instructor is required to constantly check the streaming video results, such as the position of his/her body with regard to the camera view, or the size of his/her whiteboard writing. This constant distraction makes uninterruptible teaching difficult during a synchronous e-learning session.

For this reason, computer vision has played an important role in detecting classroom events...
and transforms many probabilistic video signals into deterministic information. This information includes the instructor’s body movements, gaze, gestures, teaching objects and whiteboard writing changes. Based on this information, computer vision has made possible for automatic camera control, real-time event detection and instantaneous multimedia synchronization.

For the interested readers, please refer to Xu & Jin (2005) for more detailed discussion on the importance of these three focused e-learning areas.

**HARDWARE AND SOFTWARE CHALLENGES**

Having stated in the introduction the importance of employing computer vision technologies in e-learning applications that are synchronous, P2P based and instructor using non-computer teaching equipments, we have deduced the four main requirements associated with these types of e-learning systems, namely:

- **Bidirectional interaction (Requirement R1)**
- **Low cost hardware (Requirement R2)**
- **Higher system portability (Requirement R3)**
- **Versatile CV technology (Requirement R4)**

Following on from these requirements, in this section we will present the challenges when designing IVDA’s hardware system and computer vision-based software algorithms to address them.

**Hardware Challenges**

A typical P2P instructor’s room is depicted in Figure 1, which is also used as a testing environment for some of our experiments.

This environment setup mimics a typical P2P instructor’s room. Figure 1b is a diagrammatic illustration of the physical placements of the camera system, PC and teaching objects in relation to the instructor. In Figure 1c, we show the student’s environment. The settings are simplified to just a standard PC.

![Figure 1. A typical peer-to-peer e-learning instructor’s room](image)
In order to achieve the low-cost hardware requirement (Requirement R2), the following challenges are addressed in our hardware design:

**Hardware Portability**

A system that can easily be ported and installed from one instructor’s room to another is beneficial to the low-cost requirement, since a highly portable camera system can be shared among different users and hence more cost effective.

For this reason, in our work, the permanent fixture of any hardware in the instructor’s room is clearly prohibitive. Instead of ceiling mounting the camera(s), which is common among the other learning systems, we have placed static and PTZ camera on a tripod(s) at an eye level, shown in Figure 1a. This design has also introduced additional software issues, such that computer vision algorithms must recover quickly when accidental camera relocation occurs during a real-time session. This is discussed in software challenge section.

**Additional equipment**

We have grouped the additional equipment used in our work into reusable and e-learning specific, see Table 1.

The reusable equipment includes a consumer camcorder and its tripod, which can be reused as a household appliance or can be easily borrowed.

Our goal is to minimize the remaining e-learning specific hardware cost to be less than $200. We argue that this amount would make our system hardware accessible to most ordinary e-learners. In our project, we have used two items of e-learning specific equipment. The first one is a mechanical pan, tilt, zoom (PTZ) camera base, Eagletron PowerPod (TrackerPod, 2005), which cost $179 a piece. This item has been commercially available since 2004. The second equipment is a standard laser pointer, which cost $20 each. The total cost for these two pieces of equipment have added up to $199. In Table 1, we have itemized the prices for all the equipment used.

**Monocular Vision**

Our prototype system uses single Web cam for most of our CV detection results, apart from close-up streaming.

Two or more video capture devices provide stereovision, which can bring about many new applications into e-learning. However, many stereovision techniques are sensitive to camera position changes when a camera system is placed on a portable tripod instead of ceiling mounting. At the same time, multiple camera streams processing on single PC can also degrade the real-time performance.

**Computer Vision Software Challenges and Goals**

Given the hardware constraints listed in the previous section, the majority of our work is obviously attempting to develop and apply CV software algorithms to adapt to these hardware challenges, while achieving the four requirements stated in the beginning of this section.

**Table 1. The hardware cost in USD**

<table>
<thead>
<tr>
<th>Standard PC hardware</th>
<th>Reusable Hardware (Household Appliances)</th>
<th>E-learning Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard function</td>
<td>1 Laptop ($1,299) 1 Logitech web cam ($89)</td>
<td>1 Sony Camcorder ($599) 1 Camera Tripod ($20)</td>
</tr>
<tr>
<td>Total:</td>
<td>$1,388</td>
<td>$619</td>
</tr>
</tbody>
</table>

271
Methodology for Student-to-Instructor Communication

In order to facilitate accurate bidirectional interaction (Requirement R1), we have provided student-to-instructor communication, allowing the student to interact with the instructor’s environment by controlling the PTZ camera remotely. As a result, the student is able to view a close-up writing on the whiteboard or a teaching object without interrupting the instructor.

Automation, Intelligence and Versatilities in Instructor’s Room

In terms of vision versatilities (Requirement R4), we have developed several functions to facilitate intelligence and automations in detecting instructional video events.

IVDA is primarily for research purpose instead of product development. Therefore, instead of designing a system that can perform a quantitative set of existing computer vision functions, we are focused on incorporating state-of-the-art CV algorithms, which have not yet been applied extensively to e-learning. These algorithms are also supporting non-computer teaching as much as possible (Requirement R3).

Adaptive to Inexpensive Processors

We stated previously that many P2P instructors are home-based. For this reason, when we design the software system, we aim to customize computational computer vision algorithms to suit the user’s PC processor, instead of making the assumption that the hardware, such as fast and multi-processor computers, are always available to apply the technology (such as Shi et al., 2003). Therefore, in our experiments, we aim to achieve as much computer vision throughput as possible, without significant degradation to the visual processing results.

System Portability and CV Robustness

The system portability requirement (Requirement R3) is achieved through both the hardware and software design. Because we do not permanently install the cameras to the room, the relocation of the cameras during a live session is not always preventable. Such relocation may occur unintentionally, as the instructor may accidentally touch the camera tripod, or trip over the wires as he moves around. It can also be deliberate, such that instructor may feel the initial placement of camera does not cover certain scenes, therefore he/she may reposition the camera for a better angle during a live and informal e-learning session.

These practical challenges have prevented us from employing some computer vision algorithms where re-training or recalibration is required when camera position is changed. The techniques that fall into these categories have included background subtraction and camera calibration using reference images.

Script-Based Customizing Interface

When system is ported from one application to another, instructor will have different requirements. These requirements can be formulated into rules, which govern the camera control and multimedia syntonization. In our work, we have provided the instructor the ability to program these rules through a simple scripting interface (Xu et al., 2005).

RELATED WORK

The CV based e-learning systems have started to emerge since the late 1990s, and it has been an active and growing research field since then. In this article, we will present only the analysis of the existing works where we extracted the information
on both their applications (and relevance to our work) and the CV technologies being used.

For the interested readers, please refer to our survey paper (Xu & Jin, 2005) for a more detailed summary.

Application Areas

Automatic Camera Control

The majority of the existing works in vision-based e-learning systems have focused on automatic camera control using single or multiple PTZ cameras to simulate human video shooting. The recent works are found in Onishi & Fukunaga (2004), Onishi et al. (2000a), Bianchi (2004), Wallick et al. (2004), Rui et al. (2003), Itoh et al. (2003), Ozeki et al. (2002), Ozeki et al. (2004), Kameda et al. (2000), Kameda et al. (2003) and Shimada et al. (2004).

Our work also incorporates a PTZ camera. As we have noticed that, although capturing video data from a static camera alone is enough for detecting many vision events, however, it cannot provide a sufficient streaming result for close-up writing and objects far away from the camera. However, our work differs from these approaches, where in order to satisfy our low-cost requirement, we have used a low-cost and low-precision PTZ base compared with single-purpose and expensive PTZ camera used in exiting works.

Media Synchronization Based on Event Detection

Some of the current works also focus on instantaneous multimedia synchronization based on events captured from classroom. The work that falls into this categories include Franklin (2001), Chiu et al. (2000), Kim et al. (2004), Mukhopadhyay & Smith (1999), Shi et al. (2002), Wallick et al. (2005), Shi et al. (2003) and Shimada et al. (2004).

In IVDA, automatic multimedia synchronization is also being studied where functions such as object recognition subsystem and synchronization events scripting are proposed to facilitate this functionality.

Other Applications

There are also many other applications of computer vision technologies in e-learning. Some of these examples have included biometric identification for automatic lecturer’s login (Shi et al., 2003); apply student’s gaze information for lecture feedback evaluation ((Suganuma & Inaki, 2004)) and augmented reality in e-learning (Liarokapis et al., 2002). Currently, we feel these types of work are less relevant to P2P e-learning.

Computer Vision Technologies Used

We have reviewed the computer vision technologies used in the existing e-learning systems. Note that a complete survey of this kind is difficult, since many literatures are focusing on the learning system aspect, and the CV algorithms are often lack in detail. However, some of these technologies used can be intuitively deduced from the results they achieve.

Most of the works have used human tracking for instructor and students, face tracking, hand tracking (Ozeki et al., 2002; Itoh et al., 2003; Ozeki et al., 2004), background subtraction (Heng & Tan, 2002), gesture recognition, (Flachsbart et al., 2000; Franklin, 2001) color and motion segmentation, (Shi et al., 2003; Shimada et al., 2004), and template matching (Chiu et al., 2000; Suganuma & Inaki, 2004), as well as many other computer vision and statistical analysis techniques.

Differences to Our Work

Although many factors can be learned from the existing systems, there are also a number of dif-
ferences, which prevent some of these methods from being applied for our purpose:

**CV Robustness and Portability**

We feel that some of the existing systems have not been designed with sufficient system portability, even though they have reported impressive results under certain environments.

For example, Wallick et al. (2004) used a simplistic algorithm to track the speaker, where weighted vertical columns of the video frame difference is used to locate and update the instructor’s head position. Theoretically, this method cannot achieve high robustness in tracking under environments where the instructor is not the only moving object.

Similarly, in CABLE project (Shimada et al., 2004), the authors used a three skin color area assumption by K-means to search the face and hands region, the author then use the Hough transform to detect a circle based on the fact that faces is approximately in a circular form. The system then classifies the other two regions to be “hands.” In our experiment, this method has poor robustness when instructor is wearing cloth that has color similar to the skin.

Ozeki et al. (2004) have used a skin color model for Asian skin types, since their work is focused only on Japanese TV-learning broadcasting.

We argue that despite the high performance in their local environments, much of the system adaptability and portability are in question when CV algorithms in these works are ported from one P2P instructor’s room to another.

In addition, most of the current systems have used ceiling-mount camera systems, for example Shi et al. (2003), Itoh et al. (2003) and Ozeki et al. (2004). These methods hence prevent the hardware to be portable between the instructors’ rooms.

**Expensive Hardware**

Many existing applications that have used multiple computer vision algorithms were relying on expensive hardware to achieve their versatilities. An example of such system is Shi et al. (2003). This system is comprised of seven computers and eight video cameras.

In addition, every automatic camera control algorithm employed in the existing e-learning system has used commercial or industry PTZ cameras, where the average price for the commercial models is around $1,000 USD, let alone the industry models. These commercial PTZ cameras cannot be reused as a household appliance and are unlikely to be affordable by most home-based e-learners.

From our analysis, most of the current systems have given us the impression that their primary focus is to apply different hardware to suit the CV technology, whereas in our work, we are aiming to customize the CV technologies to suit a low-cost hardware implementation.

**INTERACTIVE E-LEARNING: CONTROL LOW-COST PTZ CAMERA**

During an e-learning session, there are many objects in the instructor’s room that need to be captured in close-up, such as whiteboard writing, the instructor and the teaching objects. This is commonly achieved using computer-controlled automatic pan-tilt-zoom (PTZ) cameras.

In contrast to current approaches, in our work we have used a low-cost, low-precision PTZ camera base to achieve automatic camera control in order to satisfy our low cost requirement (Requirement R2). Our contribution in this area has been a sophisticated and novel software control scheme to compensate the low-precision hardware disadvantage.
In order to also meet the requirement for accurate and effective bidirectional interaction (Requirement R1), we allow the PTZ camera to be controllable by an instructor as well as from a remote student. The control from an instructor is by drawing a virtual ellipse using a laser pointer over an object or a part of the whiteboard. Student control is by specifying the areas of viewing through computer interface with a single mouse drag operation.

PTZ Camera Hardware

Past Approaches

As described previously, the use of a single-purpose, professional PTZ camera conflicts with our low-cost requirement. For this reason, we have employed an alternative PTZ base, Eagletron (TrackerPod, 2005), which has been available since 2004. This device is cheap and can perform computer-controlled panning and tilting operations. In addition, it is designed to work with a household video camera, in which its optical zoom-in can also be software controlled.

The PTZ base costs USD$179. We have used it in combination with a video camera (Sony HC42E, USD$599). Although the overall camera system has only reduced the total cost marginally compared with a commercial single-purpose PTZ camera, we argue that our approach is far more economical to home-based e-learners. This is because the attached camera can be reused for household video shooting and easily borrowed from others.

Camera Configuration and Specification

Since we have already used a single static Webcam to achieve most of our e-learning vision results, therefore, it is natural to use an additional PTZ camera in conjunction with this static Webcam. The two cameras are placed sufficiently close at eye level as shown in Figure 2.

This approach of using a combination of static and PTZ camera is common among e-learning camera control schemes, such as the ones used in Bianchi (2004), Wallick et al. (2004) and Onishi et al. (2000b). Its appropriateness to e-learning is obvious, since all the areas of interest, including the instructor, whiteboard and teaching objects are naturally in the same part of the room facing and within the view of a static camera. When a PTZ camera is in action, it would have a narrower view, and hence it is relying on the static camera to capture the other events. We itemize our customized camera system's specification in Table 2.

Figure 2. PTZ and static camera configurations
Challenges for Controlling Low Precision Camera Base

Despite the low-cost advantage and the sufficiency in its panning and tilting ranges, the major setback from using this device is its low mechanical precision level; mainly due to the external joints between the base and the attached camera (commercial single-purpose PTZ cameras have mechanical integration internally).

Shown in Figure 3, by performing a +5 and −5 panning steps, the camera shows that it is slightly off-centre to the original square drawn.

The minimum steps (the step size) in pan and tilt is close to one degree interval compared with 0.07 degree interval in Sony EVID100 used in e-learning applications such as Wallick et al. (2004). In addition, this PTZ camera base does not send a feedback signal to a PC for its status during and after a movement. Therefore we cannot synchronize the current video frame feature easily with its mechanical operation.

Due to these factors, our challenge is to design computer-vision algorithms adapting to the base’s mechanical imprecision. In addition, these factors have forced us to perform camera control semi-passively based on the camera’s video feature rather than using 3-D coordinates.

PTZ Camera Control Algorithms

In our work, we need to zoom into the instructor, the static teaching objects placed on the wall (the wall-based objects) and the whiteboard writing.

The results of these three types of track and zoom operations are shown in Figure 4. In Figure 4a, the top screen capture shows when the instructor is first detected, and the bottom screen capture is after the instructor’s face has been automatically zoomed in. In Figure 4b, the top image shows the detected virtual ellipse drawn by an instructor using a laser pointer over a wall-based teaching object, and the bottom image is the screen capture after the camera zooms in accordingly. In Figure 4c, the images are similar to those of Figure 4b, except the object-in-interest is the whiteboard writing.
Placing the PTZ camera zoom onto a fast-moving instructor all the time is not possible with our low-precision camera base. Accurate and fast tracking of a moving person involves stereovision to estimate the 3-D locations of the person. For example, IBM’s person tracking project, (Connell et al., 2004) has used a pair of distant static cameras for wide stereo baseline triangulation to estimate the 3-D position of a person. Then, a PTZ camera can zoom onto the person in real-time. In addition, since our PTZ camera view is not consistent after movements (shown in Figure 3), even if there are two static cameras available to us (and calibrated without considering accidental camera relocation issue stated previously) and the 3-D position of the instructor is estimated correctly, the PTZ camera can not mechanically pan and tilt to the desired coordinates accurately.

Despite its hardware limitation, we have achieved robust track and zoom operation using the algorithm show in Algorithm 1.

The only difference compared with a real-time instructor tracking using precision camera (Kameda et al., 2003) is that, in our work, we have only allowed the zoom-in operation to occur when an instructor’s face is confined to a small movement for a five-second period. This time interval is to allow the PTZ camera to converge its mechanical movements, which then places instructor’s face in the center of its view prior to its zoom-in operation.

We argue that such confinement does not introduce a significant degradation in instructional video presentation. We have observed that, unlike the surveillance video tracking, during an e-learning session, the instructor’s upper body and face do not move all the time. They usually remain relatively stationary for a period longer than five seconds, particularly when the instructor is explaining something (which is also the most appropriate time to zoom into the instructor’s face). We would re-track the instructor once he/she began to move more rapidly.

The mechanical convergence subroutine used in the above algorithm is discussed in the subsequent sections. The face detection is based on the well known Haar-like features using Adaboost training (Viola & Jones, 2001) while tracking face after detection is using mean-shift color tracking (Comaniciu et al., 2003), which is also used in many parts of IVDA.
Estimating PTZ Camera Direction From Static Camera View

When the PTZ camera is in its normalized position, it contains a similar view to the static camera, as they are placed close together as shown in Figures 2 and 4. Therefore, if the face detection succeeds in the static camera, then it also means that the face can be detected in the normalized PTZ camera view. However, this approach also means that the PTZ camera’s mechanical convergence always starts from its normalized position.

In our work, we have also experimented with a method to control PTZ camera’s pan and tilt movement initially by static camera information. As a consequence, when the tracking subject becomes more stationary, PTZ camera’s mechanical convergence can start from a direction closer to the subject.

To achieve this, we have measured and recorded the sizes of detected faces in the static camera across different depths. During real-time, the size of detected face is then used as depth information for triangulation. A similar monocular tracking approach is also found in Cheng & Takatsuka (2005).

Although this method can not robustly place PTZ camera exactly into the direction of the instructor’s face, it achieves our purpose where in most of the time the instructor’s face is included in the PTZ camera view even though the face may not be at the centre. It then allows PTZ camera’s mechanical convergence procedure to begin when the static camera detects the instructor’s movements is slowed.

In Figure 5, we have shown the tracking result based on this technique. The first seven screen captures shows PTZ camera’s movement controlled by detected face sizes in static camera view. During this time, the instructor has large movement and PTZ camera is not exactly pointing into the direction of instructor’s face, but contains it within its camera view. In the last screen capture, instructor remains stationary for a five-second period to allow PTZ camera to complete its mechanical convergence.

Algorithm 1. Instructor tracking and zoom procedure

```plaintext
WHILE ()
    Detecting face in static camera
    IF (Face detected)
        Normalize PTZ camera’s position
        Track face in PTZ camera
    IF (Face is tracked)
        Call Start Pan-Tilt Mechanical convergence (by colour)
        IF (Tracked position changed is > Allowable Position Threshold)
            Restart Pan-Tilt Mechanical convergence subroutine (by colour)
        END IF
        ELSE
            Restart Instructor tracking and zooming procedure
        END IF
    END IF
END WHILE
```

While Detecting face in static camera
   IF (Face detected)
      Normalize PTZ camera’s position
      Track face in PTZ camera
   IF (Face is tracked)
      Call Start Pan-Tilt Mechanical convergence (by colour)
      IF (Tracked position changed is > Allowable Position Threshold)
         Restart Pan-Tilt Mechanical convergence subroutine (by colour)
      END IF
   ELSE
      Restart Instructor tracking and zooming procedure
   END IF
END WHILE
Rationale, Design and Implementation of a Computer Vision-Based Interactive E-Learning System

Zoom-in to Wall-Based Teaching Objects and Whiteboard Writings

Unlike face detection, which can be achieved automatically by the vision system, the zoom-in operation for static objects such as the teaching equipment and whiteboard writing requires user specification. The region can be specified by the instructor, using a laser pointer drawing a virtual ellipse over the subject, as shown in Figure 6a.

The system then detects the locations of the laser pointer virtual ellipse region. The laser pointer detection algorithm is based on a spatio-temporal training/detection using integral image features; its technical details are not discussed in this paper. The remote student can also specify the region using GUI interface to the instructional video, shown in Figure 6b.

There is a major difference between whiteboard writing and teaching object in terms of PTZ camera control. For whiteboard writing, the selected area has very similar color distributions to its neighboring regions. For this reason, we can not control the PTZ camera by tracking its color features. Instead, we have used a classic Kanade-Lucas Thomsi (KLT) tracking method by Shi & Tomasi (1994), where a set of “stable features” are used to track displacements between video frames using optical flow method. The track and zoom algorithm for static objects and whiteboard region is shown in Algorithm 2.

Implementations

PTZ Camera’s Mechanical Convergence

Before the zoom-in operation takes place on either type of object, the PTZ camera needs to pan and tilt until the selected region/object is in its centre. One important part of our contribution lies in our semi-passive, gradual mechanical...
convergence control algorithm. This algorithm allows the low-precision PTZ camera base to perform “centering” operation using the selected region/object as a reference.

When precision hardware is used, the correspondence between the camera’s angular motions in relation with the amount of changes in its camera view can be estimated using stereovision techniques. Then the centering can be controlled using triangulation of the 3-D coordinates. However, as we have pointed out before, such measure is unachievable from our imprecise hardware. Therefore, our semi-passive control algorithm is the only effective measure to achieve centering.

Tracking is performed on the selected/detected region using either mean-shift tracking (by colour) (Comaniciu et al., 2003) or KLT (optical flow). By accurately tracking the selected region while continuously updating the position of PTZ camera’s current centre view, its mechanical movements can effectively be converged, such that the selected region and the centre of camera view will overlap.
Algorithm 3. The mechanical convergence procedure

```
WHILE ()
    Update location of the object
    IF (pan and tilt are both less than 1 unit of movement) AND
        (Location of selected window is close enough to the center view) THEN
        PTZ camera is mechanically converged
        Call Zooming operation
    END IF
    Calculating its current direction towards center view
    IF (Last horizontal or vertical move is same direction towards center view)
        Current Move = same direction and length of last movement
        Update last move
    ELSE
        Current Move = Opposite direction, Length – 1 Unit
        Update last move
    END IF
    Stop sufficient time (by estimating its movement length)
END WHILE

CONCURRENTLY:
    Move camera according to "Current Move"

CONCURRENTLY: (either)
    Track selected objects (by color) at every frame
    Track selected whiteboard writing (by optical flow) at a given interval
```

Figure 7. Result of the mechanical convergence by tracking colour when

(a) the instructor is first detected
(b) pan and tilt operation is converged
(c) the instructor’s face is zoom-in close
(d) Image sequences captured from the PTZ camera, while it performs pan and tilt operations
The algorithm works in a simple harmonic motion style, such that it begins with larger movements. When it is getting close to an object or has “passed” over the object, its motion becomes slower and in a reverse direction. The algorithm is depicted in the Algorithm 3.

We recall that the PTZ base has no feedback signal. Therefore, in our work, both colour tracking and optical flow have to take place concurrently with the camera movements.

In Figures 7 and 8, we have shown the video sequence capture results using the above algorithm (by tracking colour and optical flow respectively). Notice that the initial movement is larger, and then it becomes slower and in a reverse direction when the object (in blue) “passed over” the centre view (in white). This procedure usually takes up to a maximum of five seconds before it converges.

**PTZ Camera Centering**

**By Tracking Color**

The tracking-by-colour algorithm is based on mean-shift tracking (Comaniciu et al., 2003), which can be executed in real-time and is particularly robust when a camera is moving. We have verified its tracking robustness under different PTZ camera movements, and noticed only under extremely large pan and tilt operations noticeable degradation is seen.

The initial face detection algorithm in Figure 7 is based on the Haar-like features training/detection (Lienhart & Maydt, 2002). We have also changed the detected face from a square to a rectangular region to improve its tracking performance.

**By Tracking Optical Flow**

Although mean-shift based tracking using color feature resulted in both accuracy and efficiency and is invariant to depth of the specified object, it however is not appropriate for tracking whiteboard writing for the reason stated earlier.

Therefore, for whiteboard writing, we have used optical flow to control its mechanical convergence and centering procedure. Optical flow is a common technique in computer vision. The algorithm calculates for a given point \([u_x, u_y]^{T}\) in image \(I_1\), a point \([u_x + \delta_x, u_y + \delta_y]^{T}\) in image \(I_2\) that minimizes \(\hat{e}\):

---

**Figure 8. Result of mechanical convergence by tracking optical flow**

(a) The top image shows the specified writing area; the bottom image shows the zoom-in result

(b) A series of images captured from the PTZ camera while it performs pan and tilt operations, with the optical flows shown in red
\[
\zeta(\delta_x, \delta_y) = \sum_{x=-w_x}^{u_x} \sum_{y=-w_y}^{u_y} I_1(x, y) - I_2(x + \delta_x, y + \delta_y)
\]

By using the algorithm stated in Algorithm 3, and replacing mean-shift tracking by optical flow techniques, we have obtained the results shown in Figure 8.

The red arrow indicates the direction of motion from the “stable” points between the frames. These “stable” features are generated according Shi & Tomasi (1994). The white arrow is the direction of the PTZ camera’s motion, calculated by averaging the feature displacement from areas containing similar depth to the board regions.

Automatic Switching Between Mean-Shift Tracking and Optical Flow Tracking

As stated above, we have used tracking either by mean-shift (color feature) and KLT (optical flow) for PTZ camera movement control. In order to automatically detect which of these two techniques to apply, as soon as the instructor selects a region using the laser pointer, we perform a color segmentation routine using mean-shift based segmentation (Comaniciu & Meer, 2002) (Figure 9) on the current video frame and test it against the instructor’s selection coordinates. If the selection contains a region of distinct color feature, then mean shift tracking scheme is adopted, otherwise, optical flow method is used.

In our earlier attempt, we developed such detection scheme based on the location of whiteboard (that is, optical flow is used for selection within the whiteboard). However, we have realized that such a scheme would constrain the instructor from placing teaching objects inside of the whiteboard area, which is often required.

Bidirectional Interactions: Control PTZ Camera from Remote Student

To achieve accurate bidirectional interaction (Requirement R1), in our work, apart from the

*Figure 9. Two segmentation results: the top image is the captured video frame and the bottom image shows segmented regions, with bounding rectangle indicating the co-ordinates of object with “distinctive” color*
instructor directing PTZ camera using a laser pointer, we also allow the student to override the PTZ camera controls remotely, as shown in Figure 6b. As a result, the student is able to choose a close-up view of the instructor room without disturbing the instructor.

A simple implementation analogous to student’s manual control of a security camera via a GUI interface over the Internet is insufficient and different to our approaches. There are two reasons:

1. Remote-controlling a PTZ camera manually takes time, particularly when camera base has low precision. During synchronous e-learning such manual effort diverts the student’s attention from his/her normal learning for a longer than necessary time.

2. The PTZ camera is operative in most of the time according to events occurring in the instructional video events. A student’s request to control the PTZ camera may need to be queued and processed a short time later.

Therefore, once the remote student has selected the region by drawing a rectangle over an area using GUI interface, no more action input is required from the student until the enlarged video stream is shown on student’s PC. The rest of the operations are handled by the vision system.

The same PTZ camera control algorithm stated in the above sections is also used to process the student’s request. There is an additional factor we need to consider:

Scheduling PTZ Camera Between Instructor and Student

In order to handle a situation when student is requesting the PTZ camera while it is in action, we have designed a camera control scheduling. In our current work, we have associated each PTZ camera’s operation with a minimum and normal allowable time. When a PTZ camera is in action, the student’s request to PTZ camera is queued until the minimum time associated with the current action is expired.

In our future work, we will use an XML approach for this scheduling. The syntax may also allow the instructor to specify the priorities to each PTZ camera’s operation, to decide whether the current operation can be immediately terminated or to have a shortened operation time. E-learning cinematic effects in PTZ camera may also be incorporated.

RESULTS, DISCUSSIONS AND CONCLUSIONS

In this article, we have identified and stated the importance of our focus areas in e-learning, which are synchronous, peer-to-peer based and the instructor uses non-computer teaching equipment. We have also illustrated the properties which these types e-learning systems exhibit and the challenges we face when designing a computer vision-based system to support these properties, which are accurate bidirectional interaction, low hardware cost, system portability and versatilities in vision algorithms.

We present each hardware component used in IVDA. We emphasised on their portability setup and inexpensiveness to reflect the low-cost requirement. We have described in detail the camera control algorithms used in IVDA, where we presented methods to control a low-cost, low-precision PTZ camera base. In this work, we have proposed several novel methods, including a semi-passive mechanical convergence control algorithm to encounter its mechanical imprecision. The control algorithm also makes automatic selection of colour-base or optical-flow base method as reference object tracking. We have also proposed PTZ camera control by both student and instructor, where instructor control is by specifying the area of interest using a laser pointer, and the
student does so by a computer interface. We also subsequently touched on the topics in camera scheduling. All these efforts are made to address the four properties we have stated.

There are also many other multimedia sub-components constitute IVDA. These sub-systems include the individual teaching object recognition (Xu & Jin, 2006) and teaching multimedia scripting (Xu et al., 2005). These combined efforts have enabled us achieving all the properties we have specified in the introduction.

The aim of our work in IVDA is also to setup an illustration by example, which, by presenting current challenges and solutions we have encountered, we hope to encourage more researchers to also consider these factors when designing vision-based automated e-learning systems. We hope there will be more progressive studies in this field, which in turn can accelerate the transfer of computer-vision technologies from laboratory into people’s everyday teaching and learning experiences.

Currently, most of our IVDA e-learning work is from a technical, particularly computer-vision design perspective. Although some usefulness evaluation analysis was made, we have yet to perform thorough pedagogical studies into the impact of IVDA system will have on different types of peer-to-peer e-learning instructors and students. Future work is also focused on collecting such information.

REFERENCES


ENDNOTES

1 Currently available to most Sony and Canon camera
2 For a Sony camera, we have controlled its optical zooming through LANC interface, for a Canon camera, optical zooming can be controlled via ZR command set.