Toward Robust Features for Remote Audio-Visual Classroom

Isaac Schlittenhart **Jason Winters** **Kyle Springer** **Atsushi Inoue** [*]
Eastern Washington University
Cheney, WA 99004 USA

Abstract

We present two studies on robustness of feature extractions for an remote classroom intelligent autopilot: (1) robust feature extractions and (2) a simple automated calibration of webcams. For the robust feature extractions, use of quantified vectors is studied as feature extractions of fuzzy classifiers in Perceptual State Machine, i.e. our core Computational Intelligence model for this intelligent autopilot. The simple automated calibration of devices is studied mainly for the sake of maximizing device utility. Those studies have shown promising results for actual use of this intelligent autopilot in ordinary classrooms that are not necessarily ideal for teleconference lectures.

Keywords: Remote Classroom, Autopilot, Perceptual State Machine, Fuzzy Classifiers.

Introduction

In this paper we present an improvement on robustness of an intelligent autopilot for remote audio-visual classrooms. This intelligent autopilot, that is currently under development, intelligently recognizes students in a remote classroom who need their instructor's attention. It controls various audio and visual devices, such as microphones and CCD cameras, autonomously as deemed appropriate. Currently, we incorporate a simple Computational Intelligence model, so-called Perceptual State Machine, i.e. a finite state machine with the use of fuzzy classifiers as its transition functions, and study mainly on its feature extractions in order to achieve our satisfactory performance. This paper describes our most recent progress on their robustness against lighting conditions of the remote classrooms that are often considered problematic for image processing.

Background and Motivation

Operating standard distance learning remote classrooms requires skilled operators. This often causes the costs of remote classrooms to be prohibitive for smaller institutions and may result in additional costs to student tuition and the institution. Further, high quality audio and visual devices often demand frequent calibration. This can require specialized, skilled technicians and generate yet another cost. As a consequence, real-time remote lectures are frequently considered infeasible and expensive, despite their potentials and needs.

Due to the recent growth in Internet communication, there is more availability of cost effective webcams, condensed microphones, projectors or large screens, and ordinary PCs (desktop and laptop). In addition to hardware, there has been a large increase of useful online services and methods of delivering remote content, such as remote desktop controls, video chat, conference calls, etc. Considering such availability of off-the-shelf products, we anticipate to utilize those in order to suppress the implementation costs.

Since all the products we are employing are ready to use out of the box, the main issue is device control and integration. Our goal is a robust and intelligent autopilot that utilizes fuzzy sets, so that off-the-shelf products are to be maximumly utilized while requiring very little or no calibration, i.e. virtually free maintenance. In addition, it is ideal to make the entire remote classroom system (consisting of off-the-shelf devices, PCs, software, and this intelligent autopilot) compact and portable, e.g. a few components on a cart. When such a system is available, remote lectures can be set in any standard classrooms within a matter of minutes by anyone (e.g. student assistants) without requiring extensive technical training.

Perceptual State Machine, our Computational Intelligence model, is essentially a finite state machine that makes use of fuzzy classifiers as its transition functions(Beaver and Inoue 2006). Those fuzzy classifiers map between perceptual states naturally recognized by human beings (e.g. 'no session', 'need attention' and 'in session') and inputs (i.e. features) extracted from video and audio streams, that are captured through physical sensors such as webcams and condensed microphones. All the previous studies on feature extractions have focused only on various pixel histograms in certain color ranges mostly for the sake of simplicity and real-time responses(Beaver and Inoue 2005; Moore et al. 2008). Those studies include palm pixel count and extension of palm pixel count, frame pixel differentiation, gesture recognition using hue saturation value, pixel count/position histograms, counting moving colored pixels, and locating a student via audio amplitude. These feature extractions performed at a satisfactory level thus held promise under ideal lighting conditions, e.g. no sunshine coming into

[*] E-mail: inoueatsushij@gmail.com

the room as a result of shutting off the window shades.

Anticipated Improvements

The following two anticipations are presented in this paper for improving the robustness against lighting conditions that are not necessarily ideal but are rather common in many ordinary classrooms. If those anticipations are successful, our technology advancement is very significant.

Toward robust question detection using quantified vectors. The well known pitfall in those pixel counting feature extractions concerns scalability and position shifts in the frame. Assuming the video frame was cropped into sections of interest what would happen if someone's foot or hand protruded into the cropped area of interest or the subject shifted rapidly in the frame? What if the individual in the frame had a large amount of skin exposed or their skin was a different hue than expected? The result was a potentially false question state when no question existed or the lack of detecting a question at all. Use of quantified vectors presented here potentially solves this problem and seems well suited to larger scale.

Toward robust images using color and lighting correction. Using off-the-shelf components has its benefits but also adds issues concerning quality control and calibration. We found that many of the inexpensive webcams do have automatic white balance and exposure controls but that these controls can be inadequate in various classroom lighting conditions. Two simple correction methods are discussed in this paper that likely improve the quality of video images that are to be fed to the intelligent autopilot.

Question Detection Using Quantified Vectors

The detection of any object through computer vision has been an evolutionary process. The real-time requirement aspect of this intelligent autopilot system puts some constraints on processing power. Because of this, more simplistic methodologies have been employed in detecting question states. Initially in Beaver's work (Beaver and Inoue 2005), a method was proposed by which pixels matching the grayscale color of the palm were counted. Through the use of a fuzzy classifier three states were identified: in session, no session, and question. Beaver's method worked well under ideally set-up conditions, but such pixel count methods can have difficulty with scale or computing distance from the camera. A following work(Moore et al. 2008), while still focusing on pixel counting, includes pixel location extracted from vertical and horizontal histograms of pixels in certain color ranges. Additionally, instead of using only grayscale palm color range, a skin detection using hue saturation value has been employed. This method has held promise and showed selected skin segments well enough when taken under a white balanced condition.

Although the pixel count methods and skin detection data plots have been distinct and held promise, they are plagued by stability and reliability issues when considering lighting conditions of ordinary classrooms. The premise of the intelligent autopilot system is for a simple cart, requiring little to no calibration, to be wheeled into a classroom and function. Through experimentation, we have found that white balance affects color and, more specifically, white balance varies from both internal and external light sources and from camera hardware. The variation in lighting causes the skin color detection algorithms to mistake objects in the room as skin. This had a very large impact on the ability to detect a question. Color change in image pixels also suffers from other problems that are difficult to address: variation in skin color, skin-colored objects in the room, individuals with large amounts of skin exposed, and shifts in scale and framing of individuals. Clearly, Color change in image pixels alone is not robust for recognizing perceptual states of students in remote classrooms.

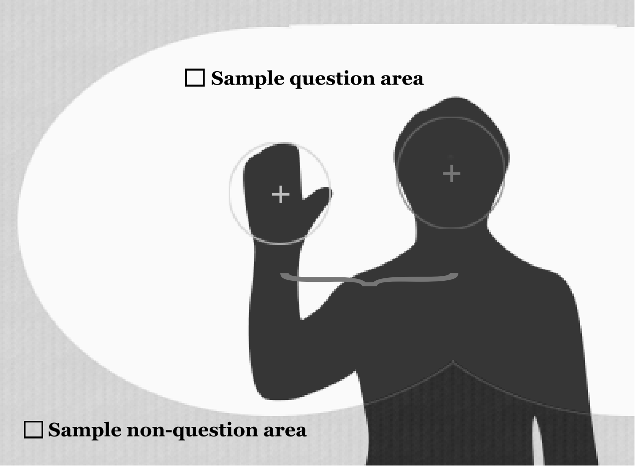

Figure 1: Sample raised hand range

With further increasing computational power and recent advancements of devices, issues of computational cost has become less significant. As a result of this, more informative methods for detecting objects in video frames are to be feasible. In this study, we propose a feature extraction method of combining hand and face position data as quantified vectors for better robustness in perceptual state recognition.

Approach

Even in live classrooms, evaluating if a student has a question is subjective. Since the likelihood of a question is not an absolute yes or an absolute no, fuzzy sets are the best suited for determining such perceptual states of a classroom.

From a human perspective, we generally and naturally analyze the location of the hand in relation to the face. Since our minds are assumed to process the position of the hand in relation to the face, we do not process this as exactly computed; but we rather simply know (i.e. perceive) if an individual has a question by observing them. Taking this into consideration, we utilize the centers of the face and the hand in such a way that a line can be drawn between them. We simply consider this as a vector. As far as its coordinate is considered, given the inherent presentation of such a vector in an image composed of rows of pixels, we use polar coordinate rather than Cartesian. In doing so, vector coordinates of the data points can be represented as x-axis distance and y-axis distance from center to center and their angles and magnitudes are inherently contained within themselves.

Since the position of the face can be described relative to the hand straightforwardly using this polar vector, the only remaining aspect that has to be addressed is how to actually recognize the hands and faces in the target image. Jones (Viola and Jones 2001) has suggested a method that has proven to be highly effective in recognizing objects given a set of training images. Other studies have shown that Haar-like classifiers proposed in their studies are superior in recognition rates per CPU cycles than many other conventional methods (Santana et al. 2008). Since Haar-like classifiers are based on trained boosted classifiers with image integrals, our current concerns such as color, white balance and skin-colored objects in the room no longer impact the feature extractions (Viola and Jones 2001).

Model

A single video frame is assumed to contain an image of the classroom. The three perceptual states of the classroom can then be outlined from a logical viewpoint as follows:

No session No faces present in the video frame.

In session Faces present in video frame but not hands (if any) in question positions.

Question Faces and raised hands present in the video frame.

First, proximities of faces and hands are presented as vectors in Cartesian coordinate such that Haar-like classifiers find the following coordinates:

- The centroid of the i-th face $F_i(x1, y1)$.
- The centroid of the i-th hand $H_i(x2, y2)$.
- The width of the face s, i.e. a scaler.

Then the quantified vector V is constructed in Cartesian coordinate such that

$$V =< x_2 - x_1,\ y_2 - y_1 > \cdot s$$

Finally this is converted into the polar coordinate such that

$$V =< r,\ \theta >$$

where $r = |V| = s \cdot \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ and $\theta = \arctan \frac{y_2 - y_1}{x_2 - x_1}$ if $x_2 - x_1 > 0$ and $\theta = \frac{\pi}{2}$ if $x_2 - x_1 = 0$. Clusters of such vectors are summarized (i.e. their histograms are generated) in order to generate fuzzy sets for the fuzzy classifiers.

Experiment and Evaluation

An open source software called OpenCV provides the necessary tools in order to train Haar-like classifiers. The experimental procedure follows:

1. Train Haar-like classifiers for hands and faces using OpenCV software. Alternatively, there are pre-configured Haar-like classifiers in OpenCV.
2. Use those trained hand Haar-like classifiers in order to detect faces and hands in a video frame.

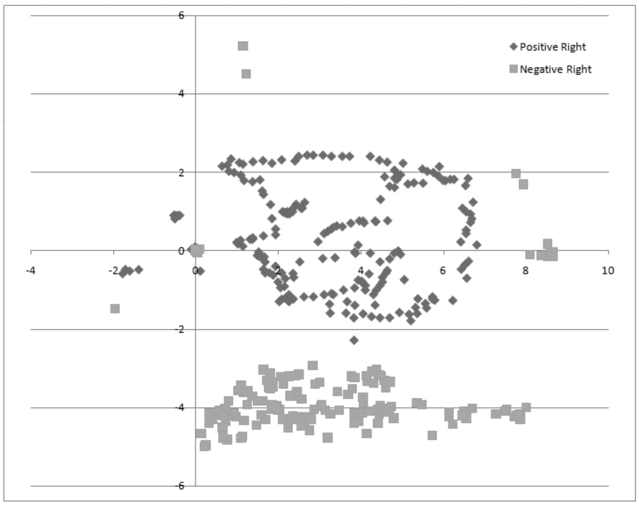

Figure 2: Sample data points

3. Collect the quantified vector in order to identify fuzzy classifiers for the perceptual state recognition.

In our experiment, we have used printed images of faces and hands then captured video frames of those through a webcam so that vector data points can be simulated as if actual images of actual students are captured. The pre-configured face classifiers have worked quite well in our experiment while the pre-configured hand classifiers do not. As a result, we have encountered some training overheads for those classifiers (a future work). Some results can be seen in figure 2. The y-axis displays the range that the hand can be located from the head vertically. This information combined with the x-axis showing the hand face distance horizontally creates a visual map that can be used to define fuzzy partitions. As shown, the data groupings are very distinct and the fuzzy classifiers should well be identified.

Findings

The robustness of using quantified vectors as features holds promise and seems to bypass many of the issues encountered with pixel/color methods. The distinct data sets in figure 2 show that this method is very distinct and will be able to recognize classroom states to a high degree of accuracy. However, a new set of issues is introduced. Haar-like classifiers may have trouble with detecting objects if the object is rotated slightly. This could pose a problem for both faces and hands. A solution has been suggested by (Barczak, Johnson, and Messom 2005) and further investigation must follow. Another issue is the dependencies of Haar-like classifiers upon a number of parameters including sample size, training parameters, and optimal training image size. For the time being, default parameter setting appears to be sufficient. It also appears that a larger sample of images are highly demanded for the satisfactory classifier training. If this is indeed the case, Haar-like classifiers may not be suitable for this intelligent autopilot. A further investigation is currently underway on this critical issue.

Figure 3: Image of excessive light coming into the camera

Robust Images Using Color and Lighting Correction

Despite some extreme lighting conditions such as a direct sun light coming into a classroom, sensory devices may still be able to capture images. In theory, our Computational Intelligence model is capable of recognizing states as long as the images are captured distinctively enough from noises, even if they are not in good quality for human eyes. The goal this work is to maximize the recognition performance as a result of color and light correction. The following two steps are considered to maximize the image quality coming into the system under extreme conditions:

1. Detect whether the appropriate magnitude of light is there.
2. If the magniture of light is not appropriate (i.e. excessive or insufficient), perform a color correction.

Step 1: Lighting Conditions

Lighting conditions in a room can greatly affect the quality of an image taken through a digital camera such as a web camera. If the amount of light in a room is insufficient, there is a chance that the camera will not pick up on details of objects in the room. If there is excessive lighting, chances that the picture is bleached out are very high. Two examples can be seen below of how the amount of lighting can affect the image taken through a web cam: figure 3 and figure 4. By looking at those images it is possible to see how badly the details of the image can be lost depending on the lighting conditions coming into the camera.

Figure 4: Image of insufficient light coming into the camera

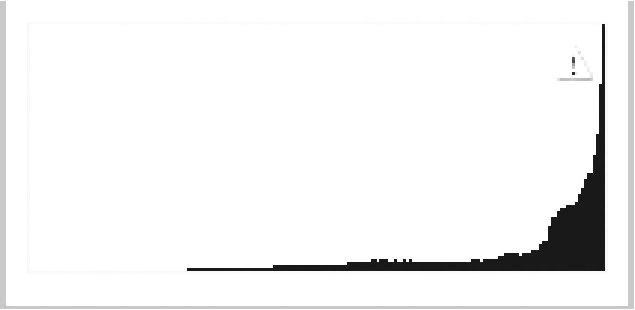

Figure 5: Histogram of excessive light coming into the camera

It is impossible to correct images that have either too much or too little light due to the fact that details in the image will be missing. This means in the system for the virtual classroom there needs to be an automated warning system to alert the user that the amount of lighting coming into the camera is not appropriate for the system. The right amount of lighting depends on the camera being used. Each camera will have a different aperture size. The smaller the aperture, the less light will enter the lens and the bigger the aperture, the more light will be brought in (Busch 2008). Since most web cameras have a fixed aperture size the lighting of the room will have to be adjusted to meet the needs of the camera selected for the virtual classroom. Traditional methods to detect the correct amount of lighting for an image taken are the use of histograms (Busch 2008). If most of the weight of the histogram falls on the left side this means that there is not enough light in the room to ensure details. If the histogram has most of the weight on the right side, this indicates there is too much lighting for the room. See 5 and 6 for examples of the histograms for the images above.

Once a histogram of an image is made it is possible to automate the process of having the system alert the user of improper lighting conditions for the lens being used inside the classroom. OpenCV has several tools to build histograms of images (Gary Bradski 2008). These histograms can be built and analyzed in real time. The data in the histograms can be placed into buckets. Each bucket will contain the count of a certain color. If there is too much black in the image, this means there is insufficient light for the camera. If there is too much white, it means there is too much light for the camera.

At this point it is unknown how much black and how much white indicates a problem with the lighting in the

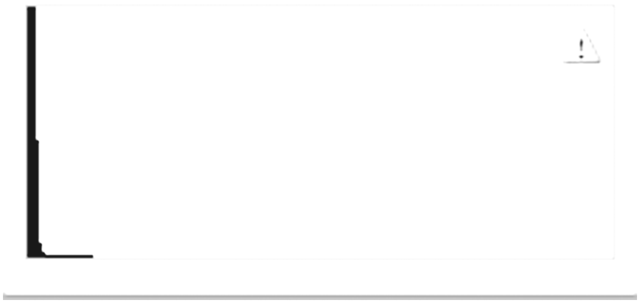

Figure 6: Histogram of insufficient light coming into the camera

system. Forty images were tested for this purpose; twenty of them were examples of insufficient light while the other twenty contained too much light. It was found that if 30 percent of them fell in the white range or black range the lighting conditions were not right for the room. If the room contains a lot of black or a lot of white to begin with the 30 percent may not be a good threshold for that room. This means that the threshold will have to be adjusted by the user if needed depending on the objects in the room.

Result 1: Lighting Conditions

To test the system 30 images were used; ten each of excessive, insufficient, and reasonable lighting. Each image was fed into the system to see if it detected the type of the input image. At this stage, ”reasonable” lighting does not mean perfect lighting. The lighting in the room just needs to be the right amount for the system to work reasonably well. Of the images tested, all the poor images were detected as ”reasonable”. However, out of the good images, two were detected as poor lighting conditions. Given that the purpose of this part of the system is just to alert the user that there could be a problem, this rate of false positives is acceptable.

Step 2: Color Correction

The next step in improving the image is white balancing the image – color correction. In any image taken the light in the scene affects the color of the image. The only type of light that is not affected is white light (light that comes from the sun). All artificial light sources contain color known as temperature (Busch 2008). The temperature changes the color of the objects in the scene. The goal of white balance is to take the color out of the objects in the scene and restore the color of the objects as if they appeared in natural white light.

While most decent image processing software packages can white balance an image, the system for the virtual classroom needs to do it in real time since the lighting conditions of the room may change while the system is in session by various lights being turned on or off. Another factor to consider is that two different cameras (even the same model) might capture images of different color values (Gary Bradski 2008). The goal of the experiment is to take an image of a known color value, see how the image is changed by the lighting conditions of the room, and correct the values of the image of the room by the values changed on the target.

The target for the test was a piece of white foam board picked up at an arts and craft store. Given that white reflects all colors, and we are looking for the color change of the target and not the color of the target itself this material seemed to be a reasonable choice (Serway 1996). Next, two cameras were tested to see if they picked up different color values. This was done by putting three colored papers (green, red and blue) in front of each camera and recording the values they saw. In each instance, camera number 2 recorded two values more red than camera number 1. It is possible to calibrate the system by adding two values to the red in camera number 1 or by subtracting two values from camera number 2.

The next step is to set up the cameras to get input into the system. One camera is always going to point at the target; the other camera is always going to point into the room. The captured image will then be color-corrected in accordance with the color change caused by the temperature of the lighting used in the scene.

To summarize:

- There are two cameras.
- The color difference of the two cameras is tested and corrected.
- One camera points at a known color target.
- One camera points into the room (audience-facing).
- The color difference between the known color of the target and the color detected in the room is used to provide color correction for the audience-facing camera.

Result 2: Color Correction

Image color correction is subjective and based on human perspective. We feel that the results of our color correction method show promise. The corrected images' color was greatly improved. The brown tint of the images caused by the temperature of the lighting disappeared and the images looked more like they were taken in natural sunlight.

Integrated Test Result

We propose that using color-corrected images likely produces better results than previously tested methods for the system. In the original study it has been shown it is possible to count the number of pixels of skin tone in the image to see if there is a hand raised or not (Moore et al. 2008). One problem is that if the temperature of the lighting of the room is not properly white balanced, objects in the room such as tables and clothing possibly looks like flesh tones to the system. During the original test it has been found that the best accuracy of the system is 86 percent. Our test images have showed an accuracy of 85 percent with good lighting to the room. With poor lighting that accuracy dropped down to about 64 percent, which indicates the impact proper lighting has on the system. After having the system correct the images, the good images have not changed but the poor images have had 73 percent accuracy. This means by having the system improve the images it is possible to increase the accuracy of the system by correcting the lighting conditions of the room.

Conclusion

In this paper, two studies on robustness of feature extractions from images are presented. They are necessary in order to achieve the goal of our intelligent autopilot to be placed in ordinary classrooms instead of ideally set-up experimental environments for many image processing works.

Use of quantified vectors that are generated by Haar-like classifiers provides a more robust feature extractions that is virtually immune to many encountered issues about scaling and positioning of objects in classroom images. Future works should easily allow for multiple individuals in each frame since a hand could be mapped to the nearest face while generating the quantified vectors, whereas multiple individuals residing in one frame posed a large problem for methods based on change in colors of pixels. Quantified vectors as robust features would potentially increase the accuracy when multiple individuals are present.

Furthermore, it has been shown that it is possible to improve the video quality through color and lighting correction. This simple method provides an automated solution to a problem commonly encountered in the ordinary classrooms that are not necessarily opted for teleconference lectures. This should bring us a feasibility that the system may be used in any ordinary classrooms (as long as a net connection is available) and that the state recognition may achieve a satisfactory level.

The next step of this intelligent autopilot is mainly about implementation and system integration for the first complete prototype.

Acknowledgment

This research was conducted as a 10-week course work of CSCD581 Computational Intelligence at EWU in Winter 2009. The authors would like to thank James Lamphere for an additional system administration support, Brian Kamp and his students for looking at bright lights while waving at the camera, and Emily Schlittenhart for her help editing and proofing this paper.

References

Barczak, A. L. C.; Johnson, M. J.; and Messom, C. H. 2005. Realtime computation of haar-like features at generic angles for detection algorithms. In *Research Letters in the Information and Mathematical Sciences - ISSN 1175-2777*.

Beaver, I., and Inoue, A. 2005. Perceptual Recognition of States in Remote Classrooms. In *Proceedings of International Conference of North America Fuzzy Information Processing Society (NAFIPS05)*.

Beaver, I., and Inoue, A. 2006. Using Fuzzy Classifiers for Perceptual State Recognition. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU2006)*.

Busch, D. 2008. *Mastering digital SLR Photography*. Thomson Course Technology.

Gary Bradski, A. K. 2008. *Learning OpenCV*. OReilly, first edition.

Moore, Z. I.; Schlittenhart, I. W.; Simpson, D. M.; Sorna, C. T.; Springer, K. A.; and Inoue, A. 2008. Intelligent Autopilot for Remote Classroom: Feature Extraction. In *Proceedings of Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2008)*.

Santana, M. C.; Déniz-Suárez, O.; Antón-Canalís, L.; and Lorenzo-Navarro, J. 2008. Face and Facial Feature Detection Evaluation - Performance Evaluation of Public Domain Haar Detectors for Face and Facial Feature Detection. In Ranchordas, A., and Araújo, H., eds., *VISAPP (2)*, 167–172. INSTICC - Institute for Systems and Technologies of Information, Control and Communication.

Serway, R. A. 1996. *Physics for scientists and engineers with modern physics*. Saunders college publishing, fourth edition.

Viola, P., and Jones, M. 2001. Robust Real-time Object Detection. In *International Journal of Computer Vision*.