SYSTEMS ENGINEERING
Research Center

A Systems Engineering Approach to Micro Expression Facial Motion Capture with Structured Light

Wade A. Bruner, Tejas Chakravarthy, Kristina E. Jones, Ruben T. Kendrick, Dustin A. LaManna

***Abstract*- A team of students at Southern Methodist University (SMU) is working with the U.S. Department of Defense (USDOD) to improve the U.S. Marine Corps training program, specifically training simulations in the U.S. to help soldiers understand non-verbal communication clues such as facial ticks. Without using facial markers an actor's face is captured and in real time transferred onto a 3-D avatar. Using one scientific camera and a process called structured light, the user's image is captured using a sinusoidal wave pattern projected over the user's face which creates a depth map of the face similar to a topographical map of the user's face. an edge detection algorithm is then applied to the depth map capturing the desired facial features, creating a wire frame. This wire frame rendering of the user's face is translated onto a mesh, creating a realistic looking avatar. Then the image is rendered and displayed to the soldier.**

I. Introduction

FOR the 2010-11 academic year the U.S. Department of Defense (USDOD) presented a project to design a micro-expression facial motion capture system to a Senior Design team at Southern Methodist University (SMU). The micro-expression facial motion capture system allows for a single actor to portray and switch between multiple simulated avatars within a small time frame. This will be used for Marine training, which currently is done by using live actors [1]. The system will allow them to cut down costs and only hire one actor to play multiple parts. A systems engineering approach was taken to solve this problem.

II. BACKGROUND

One important aspect of a soldier's preparation that is often overlooked is non-combat cultural training. Cultural training can help avoid misunderstandings, give insight to disputes and can help lead to a more general acceptance in the respective region. Naturally, a large part of cultural training is facial expression recognition, which can vary greatly between regions and is paramount to a soldier's success when deployed in an unfamiliar area [1]. The Marine Corps has shown great results with their current simulation training program, and would now like to expand the program's scope by including micro-expression facial motion capture instead of the current system being used. The current system uses a multitude of hired actors representing different disciplines from the respective cultural region.

While the first-person, actor to soldier interaction is very effective in showing common micro-expressions in real-time, it is inefficient in the sheer number of actors that need to be hired in order to portray each different role. By converting these actors to simulated avatars, the same micro-expression precision can be achieved with a significantly reduced amount of actors, to the point where the limiter is the actor's ability rather than the software capability.

III. METHODOLOGY

This section steps through the methodology currently being used to implement the facial capture system. After meeting with the client, the problem statement given was to create a micro expression facial motion capture system that eliminated the use of physical markers that must be placed on the face in order to effectively track expressions generated by the actor's face [1]. The expressions tracked by the camera must be translated into an avatar that the soldier can then interact with in real time so that the simulation can seem as real to the actual environment as possible.

A. Camera Stage

The first major aspect of the system design is the camera system that is used for capturing the 3-D image that will be manipulated before being applied to an avatar. Prior solutions to capturing a 3-D image included the use of structured light and multiple cameras that can better measure the distortions of the projected pattern [2].

The first design solution was to use two stereoscopic cameras that would each capture the different facial movements and then be compared against each other to form a final image based on a combination of the measured distortions of the sinusoidal wave pattern between the two captured images. Meshing the images from the separate cameras together is computationally expensive when capturing a still image so using this process to track facial movements would be increasingly expensive and would waste resources while also delaying the system so that the requirement of real time interaction would be difficult to meet.

Manuscript received April 4, 2011. This research is supported by DoD SERC Grant H98230-08-D-0171 and US Marine Corp.

W. A. Bruner, T. Chakravarthy, K. E. Jones, R. T. Kendrick, and D. A. LaManna are undergraduate students at Southern Methodist University.

The final solution decided on was the use of one stereoscopic camera to measure the distortions in the projected sinusoidal wave pattern. While this camera system does not produce as sharp of an image, it does allow for a faster system while still producing a clean and viewable image. The amount of matrix computations is also reduced to approximately half of those required for the two camera system which allows for the conservation of memory that would be needed to store the extra matrices and allow the system to support real time interaction with the avatar. With the reduced amount of matrix computations and manipulations, the depth map of the image can be generated faster, which will allow for less delay between the actor and the avatar replicating the actor's movements.

The camera type is also an important factor in increasing efficiency because the amount of computations needed depends on the image produced by the camera. When using a basic web camera, where the format of the image is post processed, the image has to be converted to Bayer RGB format before any other actions could occur [3]. By using a scientific camera that produces images in Bayer mode, this conversion is eliminated from the overall process, which reduces the time taken to produce the depth map of the image. In order to get a clear image, the camera lens and projector lens must be vertically aligned with each other. If this does not occur, then the image and associated depth map can be distorted because the camera will capture the distortions of the wave pattern at an angle, which will disrupt the measured distortion distance. Fig. 1 shows the sinusoidal wave pattern.

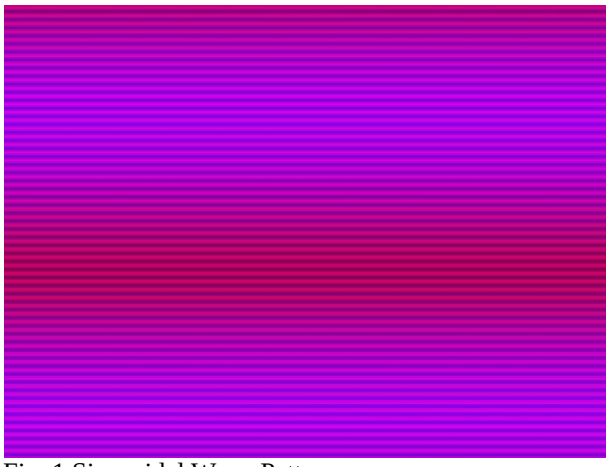

Fig. 1 Sinusoidal Wave Pattern

B. Active Illumination with Structured Light

The next step in designing and implementing the system is the incorporation of the active illumination process, more specifically the use of structured light. Active illumination using structured light is a process that involves projecting a wave pattern onto a surface so that a depth map of that surface can be generated. The depth map is generated by measuring the distortions in the projected pattern that are produced by overlaying the pattern onto the surface [4]. Fig. 2 shows the generated depth map.

Fig. 2. Depth Map

Most applications of this process are used when looking at still images such as fingerprints at the crime scene. The goal of this system is to take that process and apply it to an image that is constantly moving which requires that the pattern be continuously projected until the interaction between the actor and the soldier is completed. The pattern chosen for this process was six sinusoidal waves, three in red and three in blue, generated with Processing [5] and displayed using Open GL [6]. A green wave was not used because it would require to many calculations when the distortions of each individual pixel is measured.

A pixel contains two green sections while containing just one red and one blue section. Fig. 3 shows an example of a pixel. Measuring the distortion in the pixel using a green sinusoidal wave would require calculating the difference in depth of both green sub sections of the pixel while this process is only required for one sub section for both red and blue. Although the use of green sinusoidal waves would generate a clearer depth map, the increased amount of calculations would slow the system down and cause a greater delay in the system than was accounted for in this section of image generation [6].

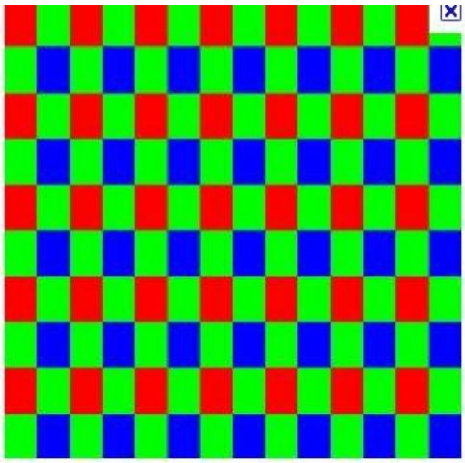

Fig. 3 RGB Pixel

These sinusoidal waves are adapted to fit the RGB color scheme and therefore only range in values from 0 to 1 instead of the traditional -1 to 1. After the wave pattern is projected onto the image, matrix calculations are done using complex numbers to convert the pixel distortion into a depth map of the image [5].

The generated depth map given on a gray scale color format with a darker pixel color symbolizing a smaller distance from the camera. White space does occur when there is an error in measuring the distortion of the wave pattern against the image. This is a small occurrence that does not affect the overall depth map so in order to save time, there is no technique used to handle this error.

C. Facial Detection Algorithm

After a depth map of the actors face has been generated, the actors face needs to be located along with the different facial features specified by the USDOD for the different cultures that they would like to represent during the training simulation. This process is accomplished using OpenCV's Haar classifier, which is implemented as a cascade of classifiers being run over the specified image. These classifiers are internally trained, although there is an external trainer also, so that they can identify the face and main facial features such as they eyes. The classifier locates the specified facial feature by dragging the search window across the image until the feature is found. The search window used is dynamically re-sized to fit the size of the feature being looked for [7]. This allows for actors with different features to use the same system where it is not specifically customized to look for specific actors facial features.

The different positions or movements that may incur over the duration of the interaction between the actor and soldier are also supported by the dynamic re-sizing of the search window. Once the search window finds the facial feature, a white box is drawn around the face and eyes to show that they have been correctly located. Although the boxes are drawn on the depth map, the Haar classifier using a picture of the image instead of the actual depth map detects the eyes and face. The location of the eyes and face are then translated to the depth map where the resulting boxes are drawn. Based on the location of the face and eyes, the next step of the system can begin where a wire frame of the actors face is generated.

IV. Future Work

A. Invisible Structured Light

Our current system uses a standard projector to overlay the sinusoidal wave pattern onto the image. This causes issues with the actor's visibility because the light that the projector uses is extremely bright and is uncomfortable for a person to look into for even a short period of time. This can cause issues in the clarity of the resulting depth map and can also cause the user to constantly shift during the capturing process, which results in a blurry image. Invisible structured light uses infrared to project these patterns instead of visible light, which would eliminate the discomfort of the actor and allow for a clearer depth map to be generated [4]. This would require an infrared projector instead of the visible light projector that is currently integrated into our system but this switch can be easily made at a later time.

B. Wire Frame and Avatar

After the position of the face and eyes has been located, a 3-D wire frame of the face will be generated using Autodesk Motionbuilder. After the wire frame has been generated, a mesh of the avatar's face can be applied over the wire frame to show the avatar with the correct facial features. This avatar will then be projected so that the soldier can interact with it directly.

V. Analysis

The system is a much more economical approach to solving this problem than the systems that already exist. For example the PhaseSpace system, which also would accomplish the USDOD goals costs six times as much as the Micro Expression Facial Motion Capture (MEFMC) System. This is shown in Table I.

Table I
Cost Evaluation of Systems

System	Cost
MEFMC	**$1730**
PhaseSpace	**$11400**
AVSoft FT2000	**$7000**

Developing the system from scratch allows full control of how the avatar is implemented and allows the system to be integrated within the USDOD system more easily. The other system's considered were either very expensive or required facial markers. Table II shows a comparison on the main requirements of not having any facial markers and the system being implemented in real time.

Table 2
Evaluation of Systems

System	Markers	Camera's Needed	Real Time
MEFMC	**No**	**1**	**Yes**
PhaseSpace	**Yes**	**4**	**Yes**
AVSoft FT2000	**No**	**4**	**Yes**

The MEFMC system developed does not require markers, uses only one camera so requires less time to set up, and is able to process the data in real time. Having to use only one

camera also reduces the complexity of the system as well as increase the portability. This makes developing the system from scratch the most economical sense as well as makes the most sense from the USDOD requirements.

VI. Conclusion

Using the method of Structured Light, the MEFMC system allows real time, high definition facial mapping of a single actor. This map will then be imported into the USDOD's training simulator and overlaid onto a digital avatar. The goal of this system is to map subtle facial features, movements, and tics needed for the correct cultural training of U.S. Marines. This system should make it immensely more efficient to train the U.S. Marine Corps.

References

[1] Kendy Vierling, Ph.D., Science & Technology Senior Analyst, USMC Training & Education Command, personal communication, August 2010.

[2] Sing Bing Kang, Jon A. Webb, C. Lawrence Zitnick, and Takeo Kanade, "A Multibaseline Stereo System with Active Illumination and Real-time Image Acquisition," *International Conference on Computer Vision*, p. 88-93, 1995.

[3] "Lenses by The Imaging Source," http://www.theimagingsource.com/downloads/choosinglenswp.en_US.pdf, 2011.

[4] J. P. McDonald, R. J. Fryer, and J. P. Siebert, "A New Approach to Active Illumination," *British Machine Vision Conference*, 1991.

[5] "Processing," http://processing.org/, 2011.

[6] Prasanna V Rangarajan, Ph.D. candidate, Department of Electrical Engineering, SMU, personal communication, January 2010.

[7] "Face Detection Using OpenCV," http://opencv.willowgarage.com/wiki/FaceDetection, 2011.