Human Motion Detection and Visual Augmentation of Chopin's Etudes

by

David Philip Kerr
B.Sc., University of Victoria, 2009

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Human Motion Detection and Visual Augmentation of Chopin's Etudes

by

David Philip Kerr
B.Sc., University of Victoria, 2009

Supervisory Committee

Dr. Brian Wyvill, Supervisor
(Department of Computer Science)

Dr. Melanie Tory, Departmental Member
(Department of Computer Science)

Supervisory Committee

Dr. Brian Wyvill, Supervisor
(Department of Computer Science)

Dr. Melanie Tory, Departmental Member
(Department of Computer Science)

ABSTRACT

Chopin's Études are difficult musical compositions for advanced piano students. Helmut Brauss, a professional pianist and educator, has created a number of videos to teach students motion patterns that will help them perfect the Études. The subtleties of motion shown in the videos are not apparently obvious to students, and in our research, we have developed four markerless based approaches to visually augment the videos: Predictive Optical Flow, Historical Optical Flow, Predictive Hand Tracking and Historical Hand Tracking. A survey of students learning the Études was conducted, and it was determined that the participants found the Historical techniques to be the most useful. No difference could be found between the usefulness of the Optical Flow and Hand Tracking augmentations.

Contents

Supervisory Committee ii

Abstract iii

Table of Contents iv

List of Tables vi

List of Figures vii

Acknowledgements viii

Dedication ix

1 Introduction 1

2 Literature Review 4

2.1 General Purpose Human Motion Detection 4

2.1.1 Initialization 5

2.1.2 Tracking 6

2.1.3 Pose estimation 6

2.1.4 Recognition 7

2.2 Human Motion Detection of Pianists 8

2.2.1 Marker Based Systems 8

2.2.2 Depth Cameras 9

2.2.3 Motion Detection Algorithms 10

2.3 Summary 11

3 Methodology 12

3.1 Optical Flow 13

3.1.1 Predictive Optical Flow 15
3.1.2 Historical Optical Flow 15
3.2 Hand Tracking 16
3.2.1 Predictive Hand Tracking 20
3.2.2 Historical Hand Tracking 20
3.3 Summary 21

4 Experiment **25**
4.1 Evaluation 25
4.1.1 Procedure 25
4.1.2 Participants 26
4.1.3 Hypotheses 27
4.2 Results 27
4.3 Discussion 33

5 Conclusion and Future Work **35**
5.1 Conclusion 35
5.2 Future Work 35

A Additional Information **37**
A.1 Survey Recruitment Script 37
A.2 Survey Questionnaire 38

Bibliography **39**

List of Tables

Table 3.1 Two dimensional histogram for figure 3.2. 14

Table 4.1 Participant demographic information. 27
Table 4.2 Participants Feedback Likert Scale ratings range from 1 (not useful) to 5 (very useful). 28
Table 4.3 Participant feedback results. 28

List of Figures

Figure 1.1 Helmut Brauss playing Chopin's Études. 2

Figure 3.1 Raw optical flow visualization, no motion. 13
Figure 3.2 Vectors displayed as red fall below the threshold when organized in a two dimensional histogram. 14
Figure 3.3 Optical flow after histogram correction, predictive flow. 15
Figure 3.4 Predictive Optical Flow visualization. 16
Figure 3.5 Historical Optical Flow visualization. 17
Figure 3.6 Historical Optical Flow visualization. 18
Figure 3.7 Hand tracking example. Original image is on the left and the corresponding binary skin frame on the right. 19
Figure 3.8 Predictive Hand Tracking visualization. 21
Figure 3.9 Predictive Hand Tracking visualization. 22
Figure 3.10Historical Hand Tracking visualization. 23
Figure 3.11Historical Hand Tracking visualization. 24

Figure 4.1 Participant feedback box plot. 29
Figure 4.2 Predictive Optical Flow visualization. 30
Figure 4.3 Historical Optical Flow Histogram. 31
Figure 4.4 Predictive Hand Tracking Histogram. 32
Figure 4.5 Historical Hand Tracking Histogram. 33

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Brian Wyvill, for guiding and encouraging me throughout my thesis work.

Michelle Mares, for mentoring, support and patience.

Helmut Brauss, for his feedback and assistance.

"Fortune is guiding our affairs better than we ourselves could have wished. Do you see over yonder, friend Sancho, thirty or forty hulking giants? I intend to do battle with them and slay them. With their spoils we shall begin to be rich for this is a righteous war and the removal of so foul a brood from off the face of the earth is a service God will bless."

Miguel de Cervantes, Don Quixote

DEDICATION

To Dulcinea del Toboso.

Chapter 1

Introduction

Video feedback of human activities has become ubiquitous in today's society. Every major sport uses video capture technology for evaluation and analysis of human motion. Athletes and coaches evaluate plays, view slow motion replays and critique the motions necessary to improve athletic performance. Musical performance requires a degree of physical dexterity similar to an athlete's but they do not receive the same degree of video feedback and analysis. A musical score guides a musician with the information to play a piece of music but it fails to offer information about the physical motions necessary to perform a musical composition. Music instructors stress the importance of proper motion patterns throughout the student's careers. But many music students, like athletes, have difficulty conceptualizing the correct motion patterns.

Athletes have found that video feedback has helped to bridge the conceptual gap so we conjecture that musicians can also benefit from this technology. Lack of motion feedback can limit a musicians potential by embedding bad habits that are difficult to correct. Repeatedly performing incorrect motions could reduce a musician's potential when dealing with difficult compositions like Chopin's Études. They can also experience injuries by repetitively performing incorrect motion patterns.

Chopin's Études are short musical compositions that are designed to be technically challenging. They require a high degree of physical dexterity and consist of complex motion patterns. Helmut Brauss is a concert pianist and master piano teacher whose area of expertise are Chopin's Études. His work, The Pianists Breviary [7] is an advanced manual for students learning the Études. To accompany the text, Professor Brauss has developed a series of video segments in which he performs motion patterns that he has designed to help pianists master the Études. Students can read the

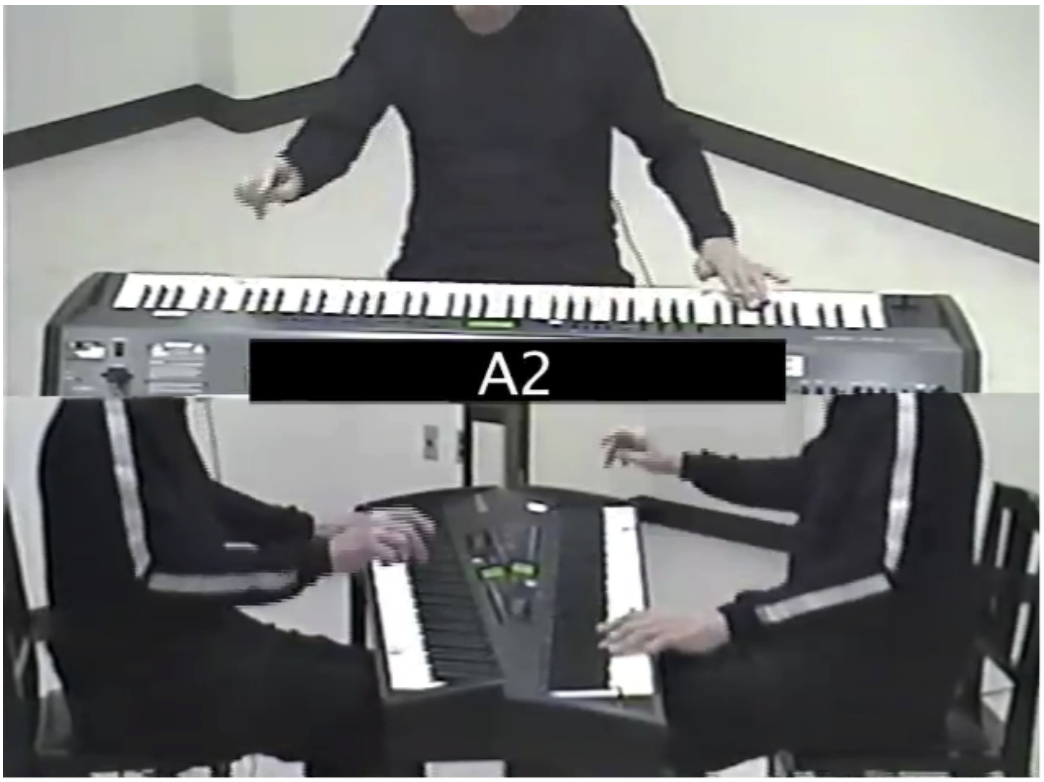

Figure 1.1: Helmut Brauss playing Chopin's Études.

manual and then reference the corresponding videos to see a visual representation of the motion patterns being performed.

The subtleties of motion in videos with no enhancement are often not immediately apparent to students. In this project we will investigate which computer vision techniques can assist students when learning Chopin's Études. Our particular research question is: What types of visual enhancements to the existing videos will increase the student's ability to comprehend the necessary motion patterns to perform Chopin's Études?

In computer vision, the results of human motion analysis offer some promising visual representations to help students comprehend physical motion. Techniques like optical flow provide a visual representation of motion in video sequences. This is a bottom up feature-based approach that is effective with a stationary camera. The ability to track specific body parts is a second area that will be investigated. This is also a bottom up approach that exploits consistencies in the video sequence to label meaningful regions.

In this research, we introduce a number of visual augmentation techniques, which

were evaluated by a group of advanced piano students. The students were shown the original and the augmented versions of the videos. A survey was then conducted to determine if there was an increased ability to comprehend the necessary motions when viewing the augmented videos.

Our main contributions are:

- Using video input, four visual augmentations of videos of the motions necessary to perform Chopin's Études have been developed.
- A study of students learning Chopin's Études was undertaken to evaluate the "usefulness" of these visualizations.
- The results show a significant difference between user preference for historically based visualizations when compared with predictive visualizations.
- The two dimensional histogram used in the optical flow computations is a compact computational representation of the motions necessary to perform each of the Études. After normalization it could be used to compare to other performers/performances.

Chapter 1 is a general introduction and statement of this projects contributions.

Chapter 2 is a literature review of general human motion detection followed by a review of human motion detection of musicians.

Chapter 3 will detail the methodology used for this project. Two techniques were used to extract the motion from the videos: optical flow and hand tracking. Each technique is visually augmented on the videos as a predictive and historical augmentation.

Chapter 4 covers the experimental validation of the project. A survey of Chopin students from the Victoria Conservatory of Music was conducted to determine the usefulness of the visual augmentations.

Chapter 5 is the conclusion and future work of the project.

Chapter 2

Literature Review

The ability to detect human motion from a video sequence is one of the fundamental aspirations of the Computer Vision community. According to Moeslund [29], human motion detection is difficult because it contains a number of ill-posed problems. The most significant difficulty in human motion detection is inferring the pose and motion of a highly articulated and self-occluding non-rigid 3D object from 2D images.

A substantial amount of research effort has been applied to the human motion detection problem but a general-purpose algorithm has yet to be found. Instead computer vision researchers apply restrictions on the domain of the problem to offer a solution that fits the characteristics of the data.

To the best of our knowledge no previous work has been done in the specific area of visually augmenting videos to teach pianists. Due to this fact this literature review will be broken into two sections. The first will discuss the literature corresponding to general human motion detection and the second will be specific to human motion detection of pianists.

2.1 General Purpose Human Motion Detection

In the field of computer vision, general-purpose human motion detection techniques are typically broken into four stages: initialization, tracking, pose estimation and recognition [29, 28].

2.1.1 Initialization

Computer vision researchers have used a number of different initialization procedures when developing human motion detection systems. The most common technique is to match a previously defined model to the subject in the image or videos being studied. To this date there is still no general-purpose method to match models to subjects. In order to achieve this task computer vision has developed a number of algorithms designed to constrain the problem including: kinematic structure, 3D shape and appearance [29].

Using prior kinematic structure has proven to be successful because accurate definitions of the number of joints, their lengths and specific degrees of freedom can be used to reduce the complexity of the matching process. Kinematic structures have the advantage of being pre-defined and can be adjusted to the particular subject being studied. A number of authors [1, 2, 33] have initialized their systems by taking kinematic structures and manually identifying joint locations in monocular images. Manually initializing models can be a time consuming process depending on the number of subjects being studied. Some researchers have attempted to automatic initialization kinematic models using video sequences. Krahnstover [23] successfully developed a system to automatically initializing an upper-body kinematic structure from a video sequence. Recent research [10, 34] of human motion detection uses multiple calibrated cameras to acquire the human subjects 3D shape. Human motion estimation techniques are used to approximate a subject shape using simple shape primitives or surfaces. Simple shape primitives include cylinders, cones, ellipsoids and super quadratics. Surface representations have included polygonal meshes and subdivision surfaces [29]. 3D shape model fitting methods are somewhat limited when dealing with real world scenes due to the variability in a subject's appearance. Most systems make a common assumption that the subject will have short hair and tight fitting clothes [29].

Researchers have discovered a significant variability between the appearances of subjects wearing different clothing. Real world examples include extremes such as subjects wearing tight fitting clothing from subjects wearing large dresses. The variability introduced is also present during the acquisition process from changes in the appearance of clothing due to motion. Appearance based initialization techniques use statistical color models of the observed image set to initialize human motion detection algorithms [29]. Body part detectors have been somewhat successful in identifying

possible joint locations using probabilistic models [27, 35]. The initialization of models that represent changes over time due to motion from hair, clothing and body shape is still an open problem [29].

2.1.2 Tracking

Surveillance applications have been the primary focus of tracking algorithms. Recent research has focused on outdoor tracking, occlusion and the detection of humans in still images. Typically tracking is broken into two stages: figure-ground segmentation and temporal correspondence.

Figure-ground segmentation separates humans from the background in an image or video sequence. Stauffer and Grimson [45] introduced the current state of the art mixture of Gaussians to perform segmentation of scenes with a stationary camera. Each pixel is represented as a Gaussian distribution over time and pixel values that do not fit into this distribution are classified as objects of interest. Motion can also be used to segment video sequences if the camera remains at a stationary position. Optical Flow or image subtraction can be used if the only difference in a video sequence is due to human motion [29].

Temporal correspondence is the process of associating the detected humans in one frame with another frame [29]. The trivial temporal correspondence case tracks one subject who is always visible in sequential frames in a video sequence. More complex algorithms and out of sequence matches are needed when dealing with multiple subjects, occlusion, subjects entering, leaving and then re-entering the scene. One advanced technique that has proven to be successful is using a correspondence matrix. Predicted objects are tracked in one direction and the measured objects are tracked in the other, the distance between the predicted and measured object can then be calculated [29].

2.1.3 Pose estimation

Pose estimation is the process of determining the underlying skeletal structure of the subject. Model based pose estimation is the process of matching a 3D model to the subject. There are two predominant techniques used to determine a subject's poise: model based and non-model based.

Model based pose estimation is the process of matching a 3D model to the subject. The majority of approaches use an analysis-by-synthesis approach to correlate

the model with the observed image [29]. Multiple and single view data sets have been employed to generate accurate representations of human subjects. Recent work with multiple cameras [10, 21] has combined deterministic or stochastic search with gradient decent for pose estimation. Constructing pose from a single view is considerably more difficult then from multiple views [29]. Additional constraints are typically used to simplify the process [8, 47]. The research community has investigated a number of approaches to monocular human motion detection including stochastic sampling [44], probabilistic approaches [25] and hierarchical kinematic models [30]. Unfortunately, the reconstruction of complex 3D human motion is still an open issue [29]. The most promising solution to this problem is in the field of learnt motion models. Using marker based data [32] as prior constraints researchers hope to achieve more reliable results than the current state of the art approaches.

Non-model based human motion detection use bottom-up approaches to detecting the pose of a subject. There are two major approaches probabilistic assemblies of parts and example-based methods. Probabilistic assemblies of parts detect the likely location of body parts and then assemble these parts to match the subject's poise [29]. Research has focused on 2D shape detection [38], SVM classifiers [41], AdaBoost [27] and appearance models [35]. Example-based methods use a database of samples to correlate the likely subjects appearance in the observed images. Sample databases can be manually labeled or data can be automatically generated from marker based image sequences. Research in this area has focused on Hidden Markov Models [6], neural networks [42] and example based approaches [18]. Example-based methods have proven to be effective when a limited training set is used. When the training set is expanded to a wider vocabulary of movements a number of ambiguities are introduced to the mapping and the reliability significantly decreases [29].

2.1.4 Recognition

The field of activity recognition can be considered old yet still immature, it is currently subject to intense investigation (see [28]). Researchers have developed a number of possible approaches to activity recognition. The applications goal and the type of data available are typically the deciding factors when researchers are deciding what approach to adapt. Moeslund [29], focuses on three categories when discussing recognition: scene interpretation, holistic recognition and action primitives.

Scene interpretation uses statistical representations that focus on distinguishing

"regular" from "irregular" activities. Typically the camera view is considered a whole and observed actions are recognized through motion [29]. Eng [13] uses features such as speed, posture, submersion index, activity index and splash index to perform surveillance on a swimming pool scene. Other methods include spatial temporal patches [4], activity trajectories [12] and nonlinear dynamical models [46].

Holistic recognition attempts to recognize human activity based on the global body dynamics without detecting individual body parts [29]. The majority of methods are silhouette or contour-based and are concerned with recognizing simple actions like running or walking. The majority of recent research has focused on scale space templates [40], hierarchical systems [39], space-time volumes [36] and temporal templates [4].

Neurobiological research has provided strong evidence that human action recognition is directly connected to the human bodies motor control system [3, 37]. The aim of action primitives are to classify a sequence of motor primitives by defining a representation and learning from demonstration. There is a large variance in the types of primitives used by the research community and the majority of work is based on motion capture data [29]. Researches have defined primitives with spatio-temporal non-linear dimension reduction [20] and non-linear differential equations [19].

2.2 Human Motion Detection of Pianists

Recently there has been an increased amount of research in the field of human motion detection of pianists. The researchers use a variety of techniques including marker-based systems, depth cameras and motion detection algorithms.

2.2.1 Marker Based Systems

Palmer [14] used a Vicon system to study pianist performing at a very fast tempi. They analyzed the finger movements of twelve skilled pianists performing a five finger melody. By placing twenty five markers on the hand and forearm of the participants they measured joint angles as they played each piece. The experiment was designed so that the tempi increased after each successive performance. They found that the pianist who were able to play the fast pieces produced keystrokes movements from the knuckle joint and only marginal movement from the fingers. In contrast the pianists who could not play the fast pieces extended the finger joints considerably.

They also found that the fast players had all joints working in the direction of the fingers movement towards the key. The slower players had some joints moving in the opposite direction of the movement towards the key.

Wristen [48] conducted a study to examine whether differences exist in the motions employed by pianists when they are sight-reading versus performing repertoire. They also plan to determine if these differences could be quantified using high-speed motion capture technologies. They used a Vicon system to capture the motion of the pianist playing two trials of a repertoire piece and two trials of a sight-reading excerpt. Angular displacements and velocities were calculated for bilateral shoulder, elbow, wrist and index finger joints. They concluded that high-speed motion capture technologies were able to quantify differences in a pianist's ability to perform sight-reading and repertoire performances. Based on the motion data they concluded that the subject's motions were less efficient in sight-reading tasks then in repertoire tasks. Unfortunately they only used one subject in their study and further investigation is necessary to determine if their findings are applicable to a generalized population.

Sakai et al [43] studied the hand motion of pianists when performing an octave chord. The subjects were required to strike two keys that are 16.7 cm apart simultaneously with the thumb and small finger. Subject's hands were covered with 26 reflective markers and the abduction angle of the thumb and small finger was measured. The researchers found that small-hand-span pianists had to abduct the thumb more then large-hand-span pianists. They concluded that this increase in abduction might cause the common pianist injury de Quervain's tenosynovitis.

2.2.2 Depth Cameras

Oka [31] uses depth cameras to determine what fingers were used to press a piano key. He defines a dictionary data set of fingering patterns consisting of a depth image, the name of the pressed key, fingering information and wrist position of the player. Unknown depth images can be matched to the dictionary to identify what finger was used to play a note. The search space is reduced in the matching process by determining the wrist position and identifying the note played through a MIDI keyboard. An evaluation on beginner piano pieces found the system to achieve a 91.6 percent accuracy rate and a process time of less then 120 milliseconds per note.

Hadjakos in [16], uses the Kinect sensor to detect a subject playing the piano from an overhead view. He develops a threshold-based technique to identify the subject's

head, shoulders, arms, hands, wrists and elbows. The head is detected as being the highest point in the depth image and the shoulders are image regions to the left and right of the head. Arms are detected by finding the boundary connected to the shoulders. Subsequently the wrists, hands and elbows are determined by shape. A marker-based system was used to evaluate the accuracy of the algorithm proposed. The researcher used a root mean squared comparison and found a error range of 3.1 to 4.9 centimetres for each body part detected.

2.2.3 Motion Detection Algorithms

Hadjakos [17] develops a system to determine which hand of a pianist plays a particular note. They use a MIDI enabled piano to record what note was played and an overhead mounted camera to capture the hand motions. The video image is processed in HSV color space and a Bayesian skin pixel detector is used to segment the subject's hands. They then compare the horizontal position of the hand to determine what hand played the particular note. They tested the system on eleven different musical compositions and through manual verification determined there method had an accuracy range of 70.6% to 99.7%.

Gorddnichy and Yogeswaran [15] develop an automatic system to detect and track pianists hands and fingers. They use a single camera, mounted from above the subject, and a MIDI keyboard to allow students and teachers to videoconference piano lessons. There system not only allows the students and teachers to communicate but it tracks what hand and finger play a specific note. They use image subtraction and deformable templates to detect the hands and an edge based crevice detector to detect individual fingers. The researchers state that their finger detection algorithm is successful fifty percent of the time.

Castellano et al [11] studied the emotional expression in musical performances. Pianists were asked to play the same musical composition with different emotionally expressive intention. Five modes were selected and labeled with emotional terms: personal, sad, allegro, serene and over expressive. The performances were filmed and analyzed with the EyesWeb Expressive Gesture Processing Library [9], a system that was previously used to analyze expressive gestures in dancers. Two motion cues were used: the quality of motion of the upper body and the velocity of head movements. Castellano [11] found that both motion cues were sensitive to emotional expression.

2.3 Summary

The field of human motion detection has been studied extensively for over 50 years. When determining what technique to apply it is best to examine the data and the expected outcome. Since we are augmenting videos hoping to help students learn advanced piano techniques we need to examine motion detection algorithms that fit into these categories. Optical Flow is an algorithm that has successful tracked feature points in a video sequence. It also offers a conceptually meaningful visual representation of motion when augmented onto a video. A second popular option is to use color based tracking to augment videos, skin color for instance can easily be segmented and tracked.

Chapter 3

Methodology

The source data for our project, was a video series of concert pianist Helmut Brauss playing what he terms as; *Motion Patterns for Chopin Etudes*, see figure 1.1. The videos contain three views of the pianist, from the front, left and right sides. The subject's head is not visible and he wears the same long sleeve dark shirt, with masking tape on his arms, for all the videos in the series. The videos have a 640x480 resolution and are filmed at 30 interlaced frames per second. The videos provide advanced pianists a visual representation of an expert playing the motion patterns devised by Helmut Brauss [7].

The aim of this project is to determine what types of visual augmentations will help students learn the motions necessary to perform Chopin's Études. Because the videos were filmed from multiple views an obvious choice would be reconstructing a three-dimensional model of the subject. Unfortunately in order to match the three camera views the intrinsic camera parameters are needed to rectify the image. It is possible to recover camera parameters by analyzing objects in the video frame similar to the work by Lee [24]. The keyboard would be an excellent choice but is only visible in the lower sections of the videos and it would be difficult to get accurate camera parameters in this position. The subject is wearing a long sleeve black shirt which makes it difficult to detect feature points that could match in different views. These limitations will significantly reduce the accuracy of any 3-D reconstruction. We decided to pursue other techniques to visually augment the videos namely optical flow and hand tracking.

3.1 Optical Flow

Because the camera position remains constant, and the only motion is that of the human subject, the Lucas-Kanade [26] optical flow technique is effective at tracking the subject's motion. The OpenCV [5] implementation was used to detect features in one frame of the video sequence and match the same features in the next frame. This process was repeated for all frames in the sequence and disparity between matched features was converted into motion vectors that can be displayed to the user.

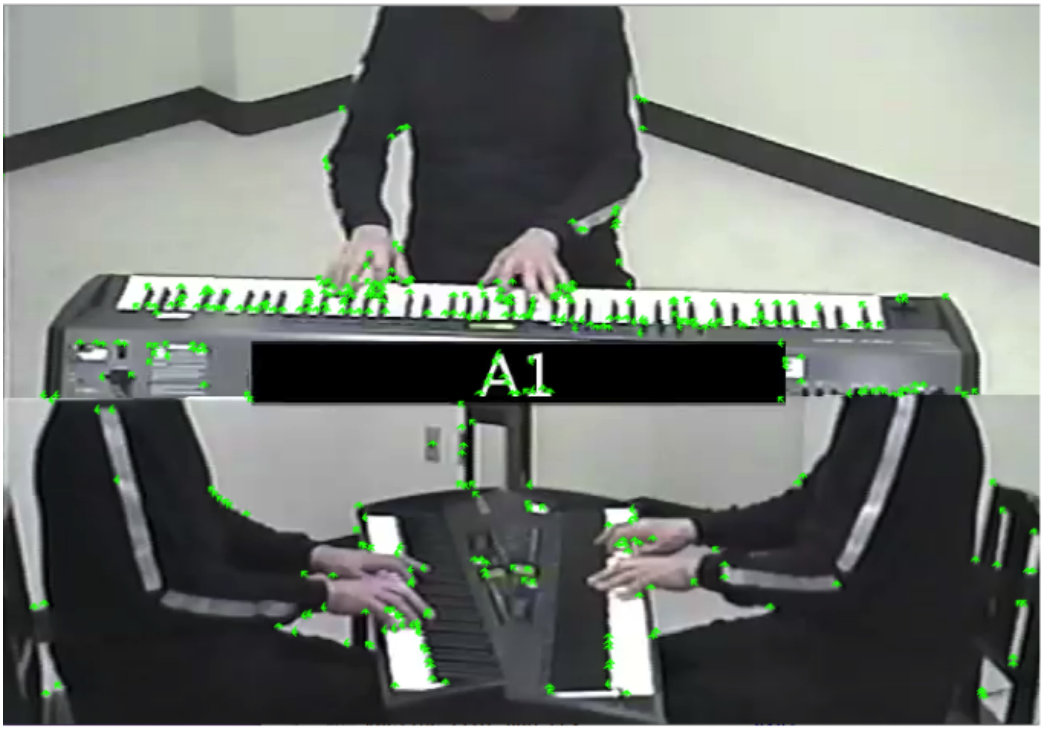

Figure 3.1: Raw optical flow visualization, no motion.

The raw optical flow values are extremely noisy due to the low quality of the videos, see figure 3.1. Features were tracked in one frame and then incorrectly matched one pixel away in the next frame. To eliminate the false positives a multi-dimensional histogram was used to bin the vectors and remove entries that fall below a threshold value, see table 3.1. Vectors are grouped based on length and angle; fifteen pixel values are used for vector length bins and twenty for the vector angle bins. This creates a two dimensional histogram of vectors that can be processed to determine the most significant motions. Because of the low video quality vector length values less then two pixels must be removed and we require that any other bin must have

Bin Length (Pixels)	0°-72°	73°-90°	91°-108°	109°-126°	127°-144°	145°-162°	163°-180°	181°-360°
1	X	X	X	X	X	X	X	X
2	0	5	21	1	7	0	2	0
3	0	0	0	0	23	0	5	0
4-15	0	0	0	0	0	0	0	0

Table 3.1: Two dimensional histogram for figure 3.2.

at least three vectors to be augmented on the video, see figure 3.2.

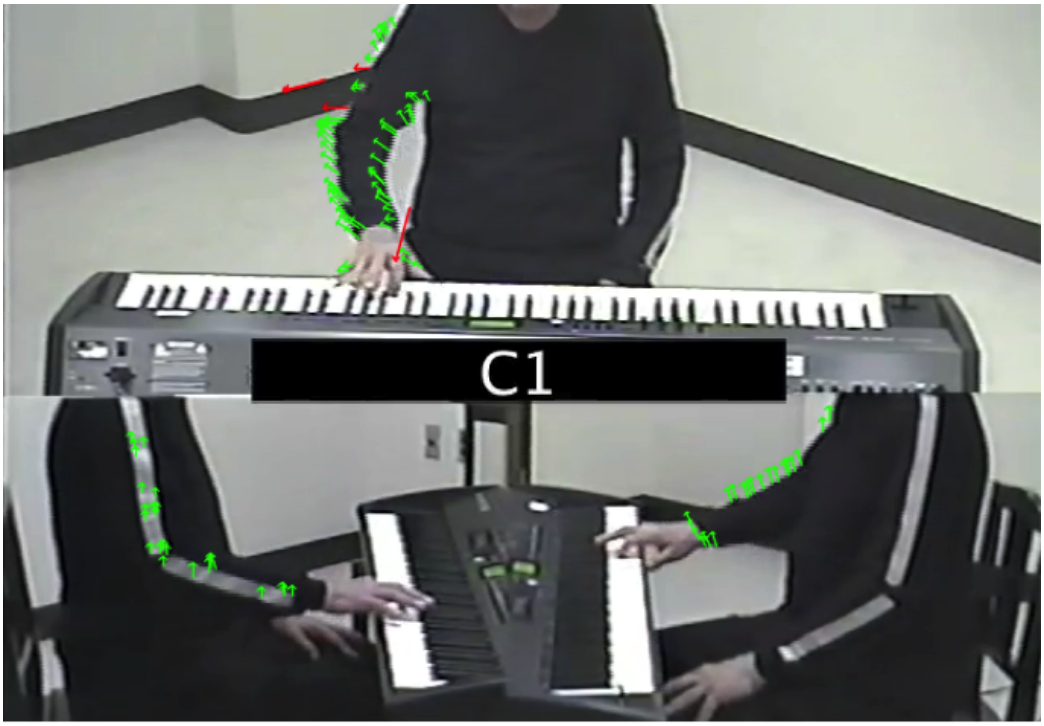

Figure 3.2: Vectors displayed as red fall below the threshold when organized in a two dimensional histogram.

The multi-dimensional histogram has a variety of potential uses because it contains a compact representation of the motion related data to perform each note of the motion patterns. This project is focused on visual feedback of the motions. Two different visualizations techniques have been developed from the two dimensional histogram data, and have been labeled predictive optical flow and a historical optical flow.

3.1.1 Predictive Optical Flow

The predictive optical flow uses the data in the corrected two-dimensional histogram to draw vectors on the subject, see 3.3. The direction of these vectors correlates with the direction of motion present in the video sequence and the length correlates with the velocity of the motion. The vectors are considered predictive because they display the motion that the subject is about to perform.

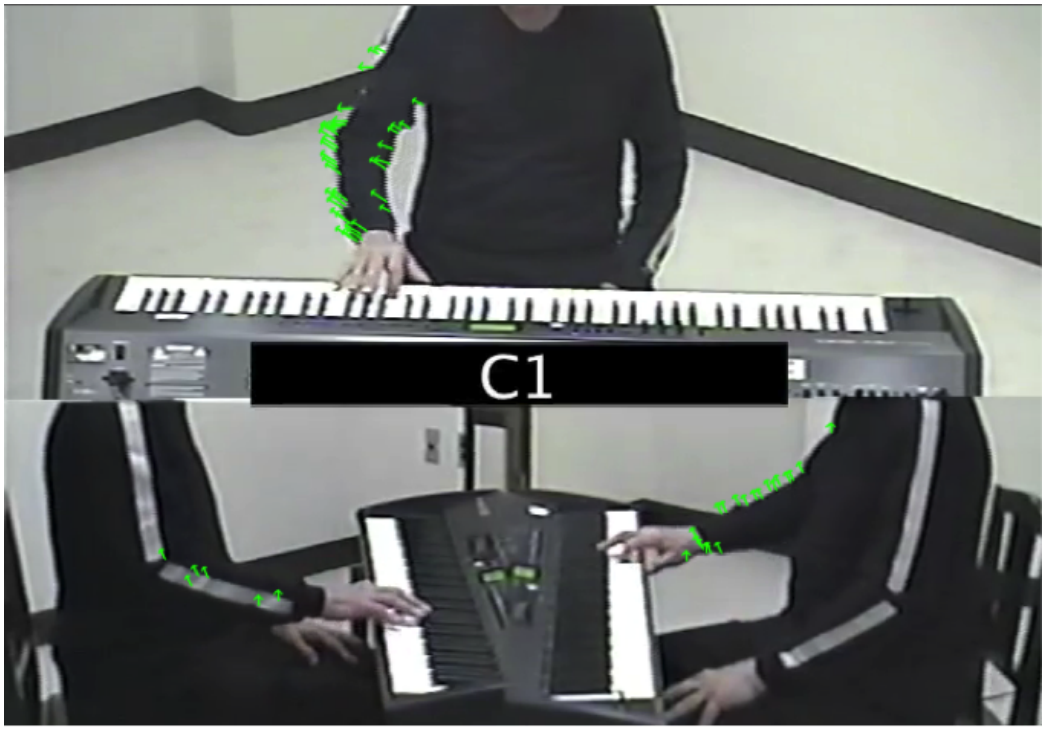

Figure 3.3: Optical flow after histogram correction, predictive flow.

Figure 3.4, shows a sequence of predictive optical flow frames where the subject is shifting his arm to the right for frames (a)-(d), paused in (e), to the left in (f) and (g) and paused again in (h). The predictive optical flow shows a visual representation of the motion in both direction and velocity.

3.1.2 Historical Optical Flow

The historical optical flow uses the data from the corrected multi-dimensional histogram to draw vectors from previous frames on the subject, see Figure 3.5. The vectors of the previous ten frames are drawn on a single frame. The direction of these vectors correlates with the direction of motion present in the video sequence

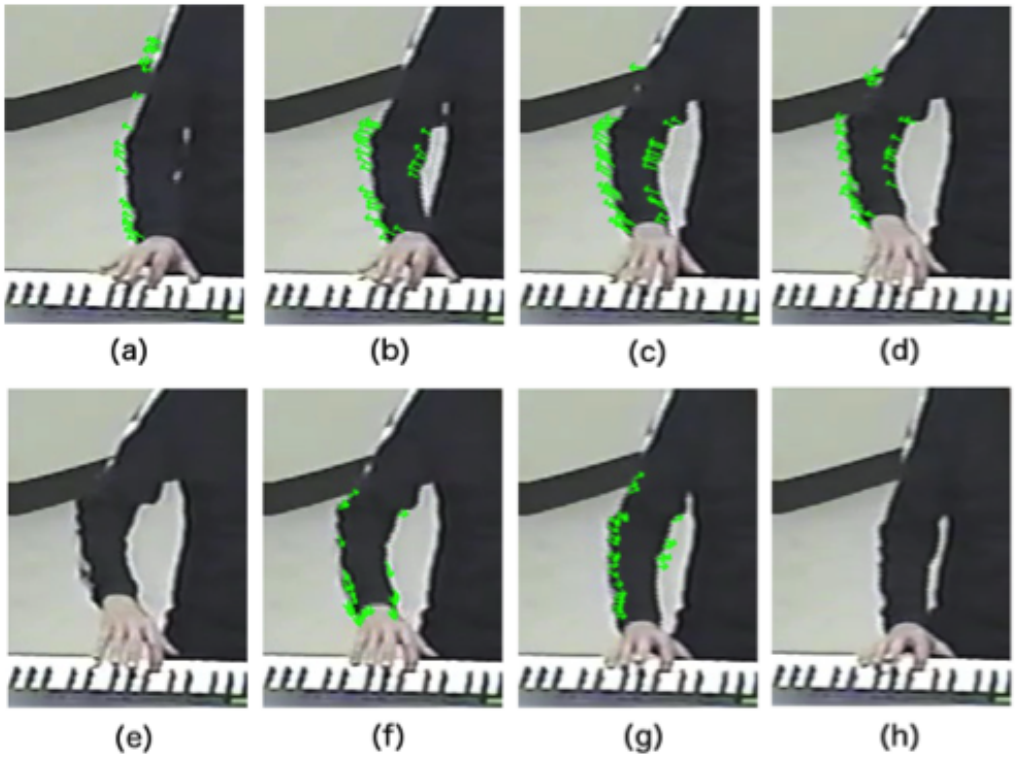

Figure 3.4: Predictive Optical Flow visualization.

and the length correlates with the velocity of the motion. The vectors are considered historical because they display the motion of the subject from previous frames.

Figure 3.6, shows a sequence of historical optical flow frames where the subject is shifting his arm to the right for frames (a)-(d), to the left in (e) and shifting his hand up and to the right in (f)-(h). The historical optical flow shows a visual representation of the motion in both direction and velocity.

3.2 Hand Tracking

Professor Brauss [7] recommends that tracking siginicant features like the head, elbows and hands. He believes these features will offer the most benefit to students learning Chopin's Études. The head would be an interesting feature to track in the video sequences because the motion relates to the mood or ambiance of the Étude [11]. Unfortunately, when the videos were created the head was removed to increase the field of view. There are some segments where the lower section of the subjects face is visible and it has been tracked when available.

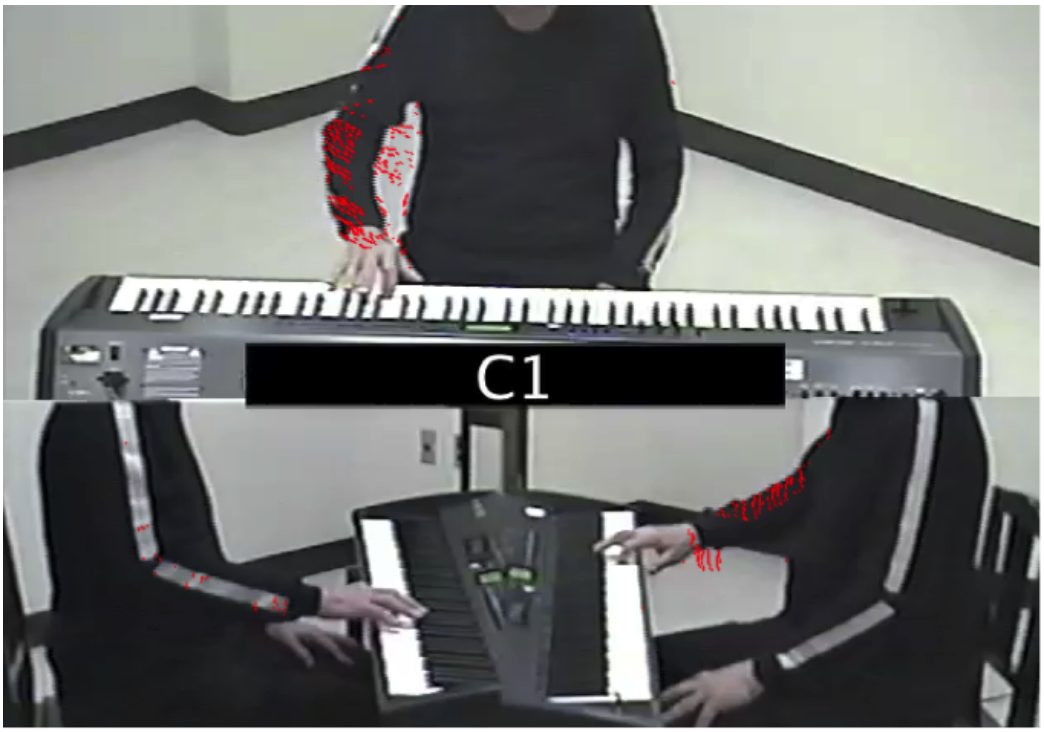

Figure 3.5: Historical Optical Flow visualization.

According to Brauss [7] , the elbow position is important for students to understand when learning the Études. Elbow position is directly related to the hand position over the keys and allows the pianist to move fluidly from note to note. Proper elbow position will also dramatically reduce injury when a pianist is practicing for many hours. None of the techniques we implemented in this research were capable of reliably tracking the elbows. The problem is finding a reliable feature close to the elbow to track throughout the videos. Because the subject is wearing a long sleeve black shirt no obvious features are available. An attempt was made to track the sleeve wrinkles near the elbow but these features are sometimes occluded by the subject's motions and not consistent. There is also an issue when the subject's elbow is inline with the dark baseboard in the background because a black object moving over a black background will not display any trackable features. It would be possible to track the elbow in the side views because the subject has marked his arms with masking tape but after talking with Brauss in was concluded that the most useful view is from the front.

Brauss [7] believes the motion of the hands is the most important feature in the video sequences. The hands are visible throughout the majority of the videos. At

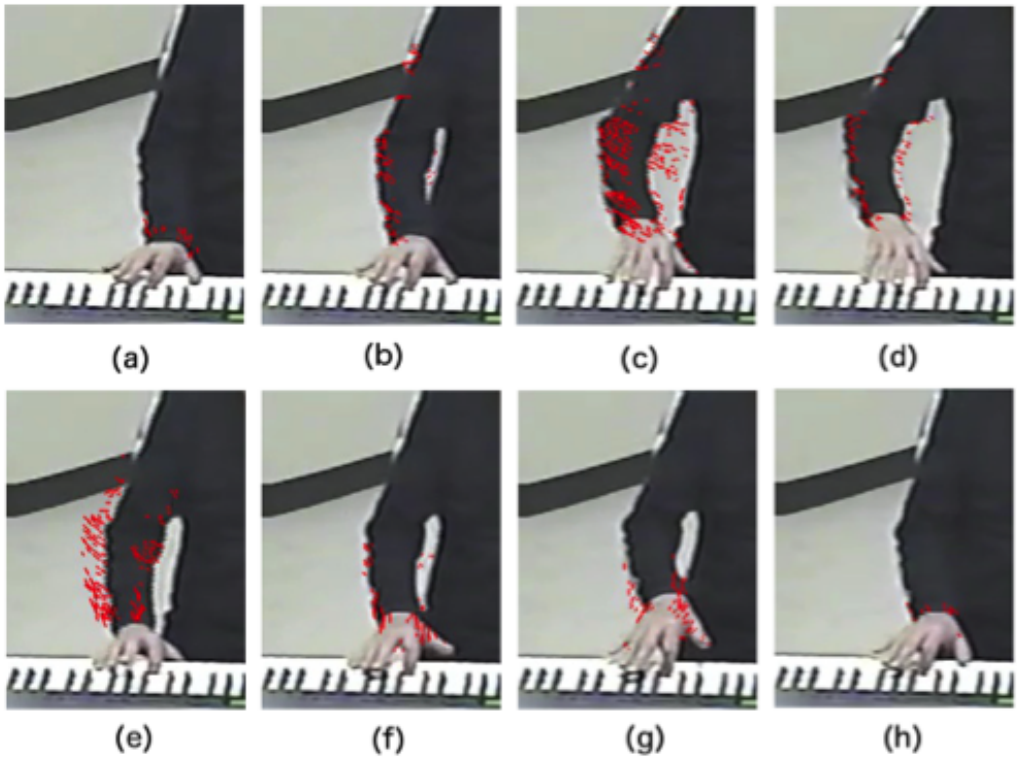

Figure 3.6: Historical Optical Flow visualization.

times the hands are occluded by the keyboard but this situation usually occurs when the hands are at rest before or after the performance. The hands also occlude each other at times in the videos. This is most likely to occur during the side views of the performance. The performer is wearing a long sleeve black shirt and the hands have a definitive boundary. Therefore, a multiple thresholding technique will easily segment the hands from the rest of the objects in the videos. Kovac et al. [22] proposed the following multiple thresholding technique for colour images with RGB values ranging from 0 to 255:

$$\begin{aligned} R > 95,\ G > 40,\ B > 20 \\ \max(R,G,B) - \min(R,G,B) < 15 \\ |R-G| > 15,\ R > G,\ R > B \end{aligned}$$

Every pixel of each frame in the video sequence was tested against these seven thresholds. If the pixel successfully passed all seven thresholds it was labelled as a

skin pixel. Using these values was moderately successful for the videos of Helmut Brauss performing Chopin's Études. Portions of the hands were tracked throughout the video sequence but there were also a number of false positive pixels. The following values were refined using a heuristic technique for the videos and used in this project:

$$\begin{aligned} R > 80,\ G > 25,\ B &> 5 \\ \max(R,G,B) - \min(R,G,B) &< 5 \\ |R-G| > 5, R > G,\ R &> B \end{aligned}$$

The OpenCv simple blob detector [5] was used to detect the centroid of the skin blobs. This function takes the binary skin frame as input and returns the x and y coordinates of the centroids of each blob in the image, see figure 3.7. In order to eliminate the small false positive blobs that were detected in the image the minimum area of a blob was set to 50 pixels. To ensure that large sections are classified as one blob the minimum distance between blobs is set to 30 pixels. This can cause the hands to be classified as one blob when they are close together but it eliminates the possibility that one hand can be labeled as two separate blobs. The position of each blobs centroid is then saved into a data structure that is indexed by the frame number in the video sequence.

Figure 3.7: Hand tracking example. Original image is on the left and the corresponding binary skin frame on the right.

The final stage of the hand-tracking algorithm correlates each blob in a particular frame with the same blob in the next frame of the videos. The predictive tracking needs to link the current blob with the same blob in the next frame and the historical tracking needs to index the same blob in the previous frame so that a line can be

drawn connecting the two. An index vector was used to link the centroids of each blob from frame to frame. The Euclidian distance between a blob in one frame was measured against the blobs in the successive frame. The blob with the minimum Euclidean distance was chose to correlate with the blob being indexed in the current frame as long as that distance was less than 25 pixels. If the minimum distance is found to be greater then 25 pixels for each blob an empty index would be assigned to that blob. Empty indexes also occur when the hands become occluded by the keyboard or when the hands occlude each other.

The data from the hand-tracking algorithm is visually augmented onto the videos with two separate visualization techniques: predictive and historical hand tracking. These visualizations are similar to the optical flow visualizations but instead of tracking all available feature points only the significantly meaningful hand points are displayed.

3.2.1 Predictive Hand Tracking

The predictive hand tracking displays the motion in the current and the next frame in the video sequence. Vectors are drawn over the hands displaying the direction and velocity of the motion, see Figure 3.8.

Figure 3.9 shows a sequence of the right hand in the predictive flow visualization. The hand is stationary in frame (a) and moving up and to the right in frames (b)-(d). Frames (b)-(d) illustrate a velocity change in the hand motion. The velocity of the hand motion is equivalent to the length of the feature vector. For example, the vector is longer in frame (c) than it is in frames (b) and (d). This gives a visual representation of the velocity changes. The hand is moving the right in frame (e) and decreasing in velocity as it moved down in frames (f) (h) because the length of the velocity vector is decreasing.

3.2.2 Historical Hand Tracking

The historical hand tracking draws lines connecting the hand position in the previous twenty frames of the video, see Figure 3.10. The users view a historical augmentation of the hands and the visualization conveys the motions necessary to play the Étude.

Figure 3.11, shows a sequence of the right hand playing one of Chopin's Études. The displayed sequence shows frames that are approximately five frames from the previous. The hand is stationary in frame (a) and moving up and to the right in

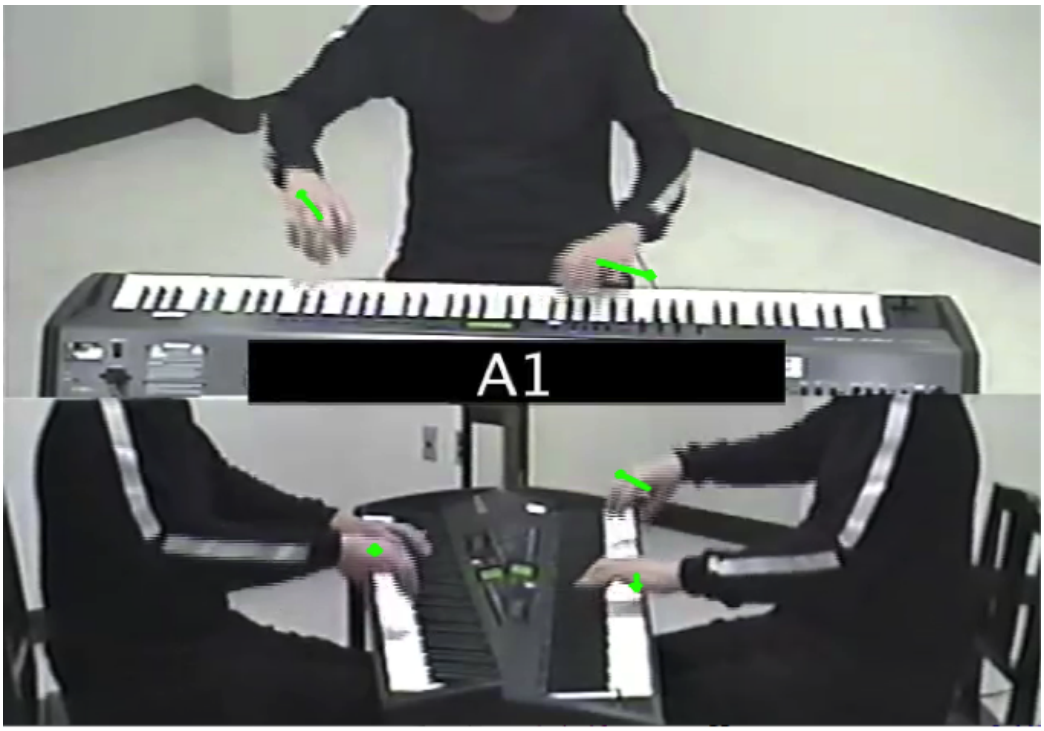

Figure 3.8: Predictive Hand Tracking visualization.

frames (b) through (d). The hand has reached its apex in frame (e) and is moving down to strike the key in frames (f) through (h).

3.3 Summary

An optical flow based algorithm has been presented that reliably extracts and verifies the significance of motion vectors of a subject playing Chonin's Études. We have also introduced a color based tracking algorithm that matches color blobs in subsequent frames. Each algorithm has been augmented with a predictive and historical visualization. Using these techniques it is possible to automatically augment all the video of Helmut Brauss performing Chopin's Études. These videos can now be presented to an audience of advanced pianists that are at various stages of learning Chopin's Études. We hope to verify the "usefulness" of the visual augmentations by soliciting feedback from these pianists.

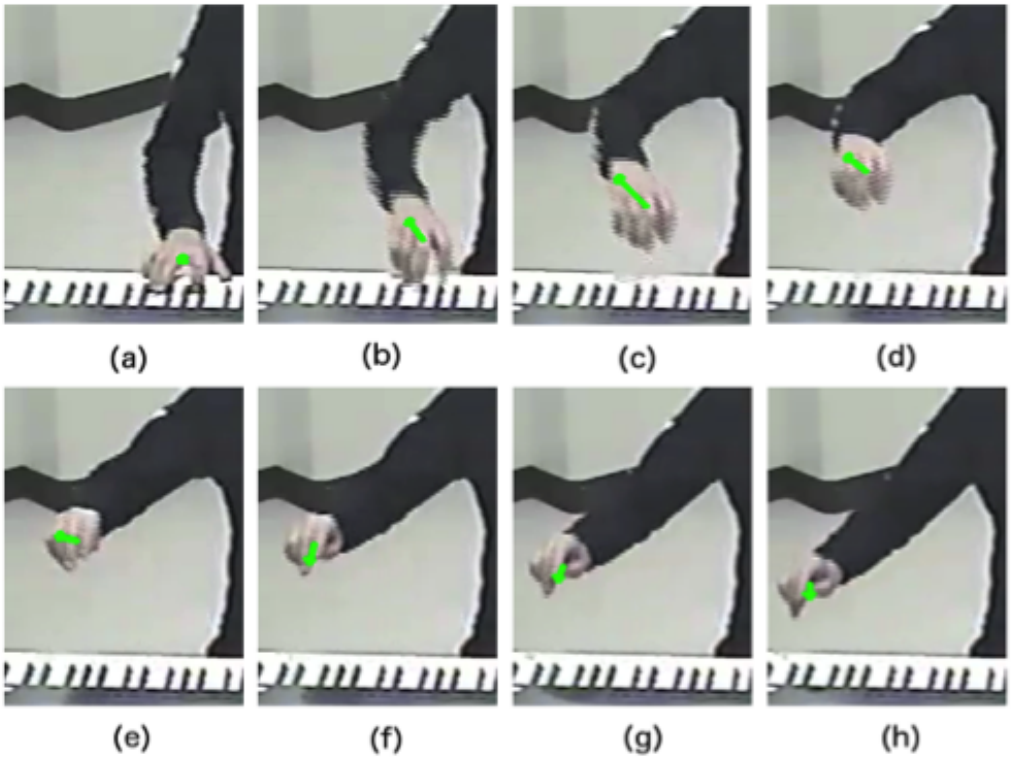

Figure 3.9: Predictive Hand Tracking visualization.

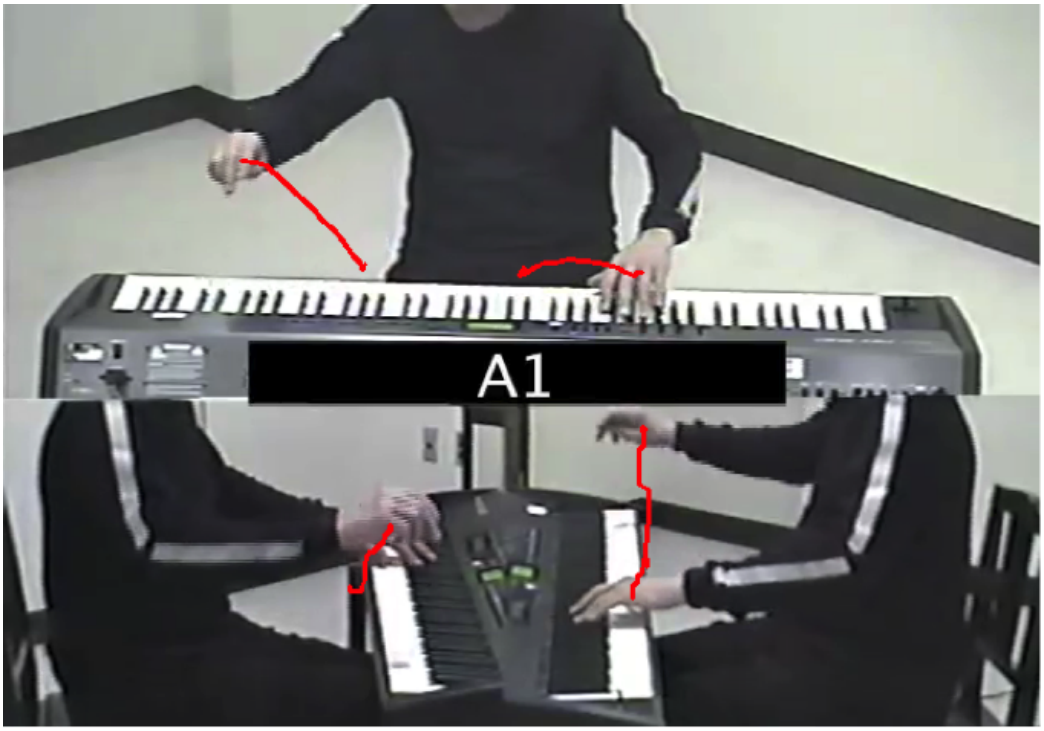

Figure 3.10: Historical Hand Tracking visualization.

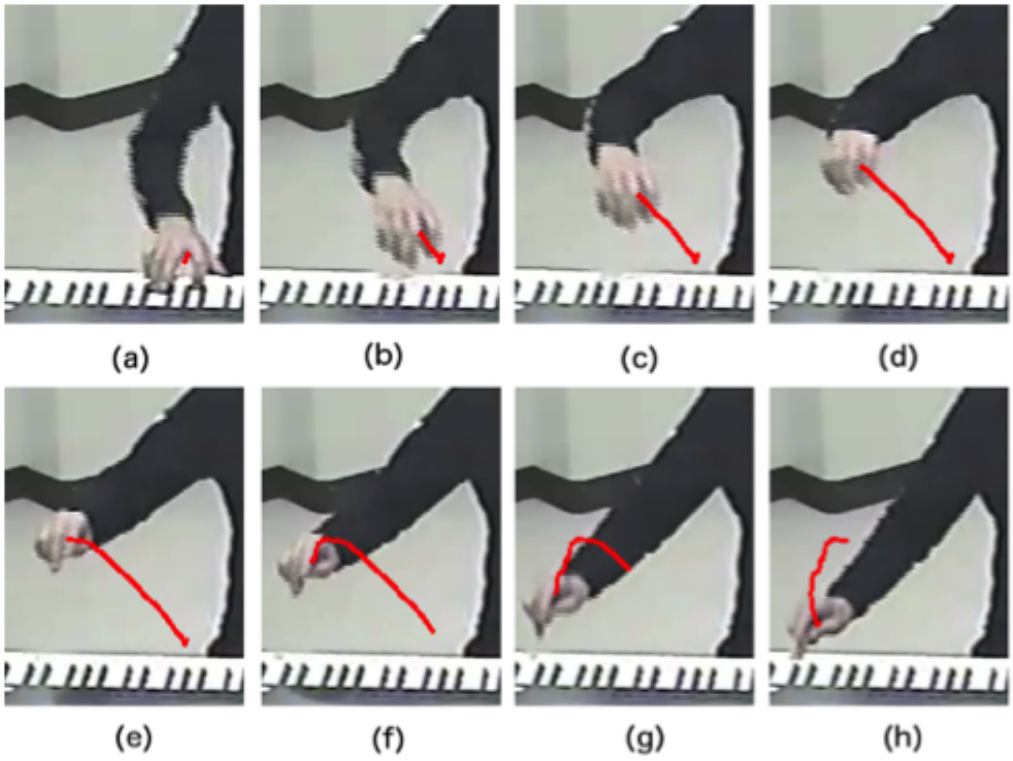

Figure 3.11: Historical Hand Tracking visualization.

Chapter 4

Experiment

4.1 Evaluation

We have developed four separate techniques to visually augment videos of a pianist performing motion patters for Chopin's Études. We have shown that the algorithms used in this work are effective at extracting human motion from the video sequences, but the essential question is whether the visualizations are useful to students learning Chopin's Études? What visualizations do the students find to be most useful? Do they prefer the optical flow or the hand tracking visualizations? Do students prefer the predictive or the historical visualizations?

4.1.1 Procedure

In the summer of 2013, concert pianist and piano instructor at the University of Victoria, Michelle Mares, taught a two-week class to advanced piano students learning Chopin's Études. The class met for one hour each day to discuss the technical and musical challenges presented by each of the etudes. During class time, the instructor relied on live demonstrations as the primary instructional aid. Approximately half way through the course, the videos were introduced. A sampling of the videos were shown during class, and then each student was provided with a CD of the videos to study at home. The CD contained the original videos (made by Helmut Brauss for his book, The Pianists Breviary [7]) plus the four types of augmented videos. It was recommended to students that they view the videos in slow motion and at full speed. Students had approximately one week to study the videos.

This research study was not mandatory for the students and declining to par-

ticipate did not affect their evaluation in the course. At the end of the course, participating students were given a questionnaire that asked their date of birth, years of piano experience, months of Chopin experience and how many hours a week they practiced piano. The questionnaire then asked them to rate each visual augmentation on a one to five point Likert scale, one being not useful and five being very useful. The questionnaire also asked them to provide free-form comments about each visual augmentation. The questionnaire was handed out in class but completed outside of class to avoid taking classroom time.

4.1.2 Participants

A total of eight participants from the Victoria Conservatory of Music elected to partake in the user study. We have enumerated each participant so they can be referred to anonymously. Table 4.1 displays the participants' demographic information. This data was used to verify whether the participant was suitable to partake in the study.

The participants' age range was quite extreme. There were three young participants, four veteran participants and only one middle-aged participant. Participants' piano experience correlates with age, since most advanced pianists start when they are young. Only three participants had any significant Chopin experience. The rest had no Chopin experience, except participant seven who reported having two months of experience. All participants spent a great deal of time practicing per week. There were four participants who practiced over fifteen hours per week, three who practiced between ten to fifteen hours, and one who practiced six hours per week. All participants met our minimum requirement of playing piano for at least five years and practicing over five hours per week.

Note that it is very difficult to find participants who are qualified to play Chopin's Études, as these are very challenging pieces even for advanced pianists. Therefore, the number of subjects that are qualified to give feedback on the visual augmentations is limited. Although eight participants is low for a typical user study, finding this many participants simultaneously represents a significant achievement.

Participants	Age (Range)	Piano Experience (Years)	Chopin Experience (Months)	Practice Weekly (Hours)
1	15-20	12	0	18
2	15-20	11	0	18
3	20 -25	10	0	18
4	35-40	20	24	11
5	50+	40	0	6
6	50+	50	75	20
7	50+	52	2	14
8	50+	67	24	12

Table 4.1: Participant demographic information.

4.1.3 Hypotheses

- Q1 - Will the participants find the hand tracking augmentations more useful than the optical flow augmentations?
- Q2 - Will the participants find the historical augmentations more useful than the predictive augmentations?

Our hypotheses were developed based on discussions with Chopin experts prior to the study. Expert feedback suggested that students would gain a higher-level understanding about the motion by viewing semantically meaningful regions like the hands as opposed to seeing all the visible motion (H1). Experts also suggested that a historical augmentation of the motion could be more beneficial to the students than the predictive augmentation (H2) because it conveys more detailed information about the path and shape of the movement.

4.2 Results

A total of eight participants from the Victoria Conservatory of Music elected to partake in the user study. We have enumerated each participant so they can be referred to anonymously. Table 4.1 displays the participant's demographic information. This data is used to verify if the participant is suitable to partake in the study. The participant's age range is quite extreme. There are three young participants, four veteran participants and only one middle-aged participant.

Participants	Predictive Optical Flow	Historical Optical Flow	Predictive Hand Tracking	Historical Hand Tracking	Totals
1	2	4	1	5	12
2	3	5	5	4	17
3	2	4	4	5	15
4	4	4	2	2	12
5	2	3	2	3	10
6	3	2	3	4	12
7	1	4	1	4	10
8	2	4	4	3	13

Table 4.2: Participants Feedback Likert Scale ratings range from 1 (not useful) to 5 (very useful).

Method	Predictive Optical Flow	Historical Optical Flow	Predictive Hand Tracking	Historical Hand Tracking
Mean	2.38	3.75	2.75	3.75
Median	2	4	2.5	4
Mode	2	4	1	4
STD	0.92	0.89	1.49	1.07

Table 4.3: Participant feedback results.

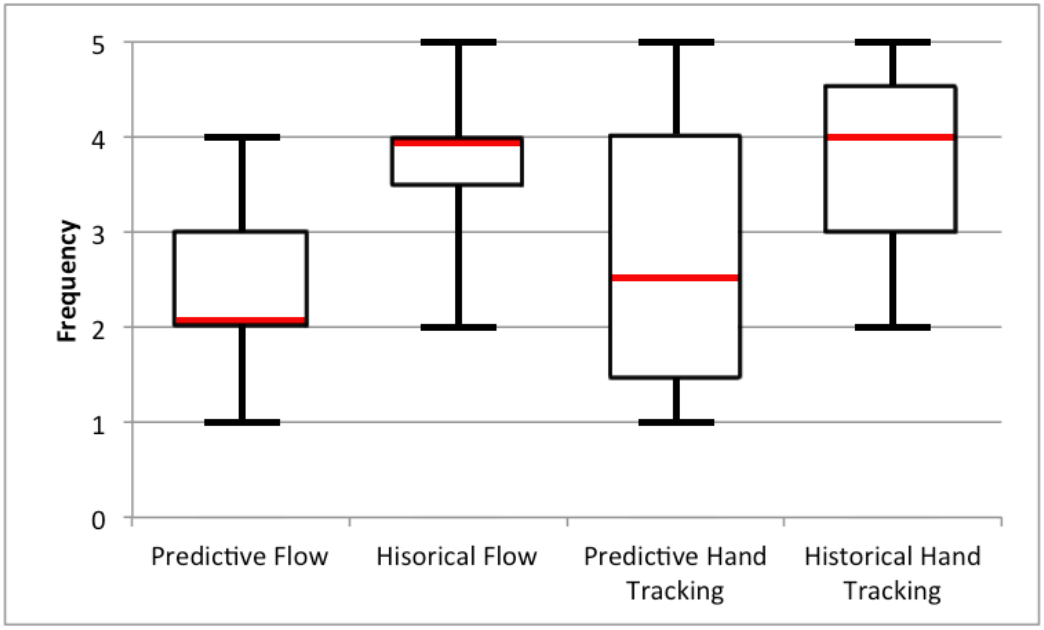

Figure 4.1: Participant feedback box plot.

The participant's piano experience correlates with the participant's age, since most advanced pianists start when they are young. Only three participants had any significant Chopin experience. The rest had no Chopin experience except participant number seven that reported having two months of experience. All participants spend a great deal of time practicing per week. There are four participants that practice over fifteen hours per week; three that practice between ten to fifteen hours and one that practices six hours per week. All participants meet the minimum requirement of playing piano for at least five years and practicing over five hours per week.

Table 4.2 displays the participants feedback for the visual augmentations. The feedback is based on the Likert scale ratings ranging from 1 (not useful) to 5 (very useful). Table 4.3 displays the mean, median, mode and standard deviation results for each visual augmentation. The ratings are based on a Likert scale ranging from one (not useful) to five (very useful). Figure 4.1 compares the four techniques in a boxplot. As shown in the figure and table, the historical techniques generally received higher scores than the predictive ones.

We evaluated the results statistically using a series of planned comparisons. Specifically, we wanted to compare (1) Optical Flow and Hand Tracking techniques for each of the predictive and historical methods (to assess H1) and (2) predictive and historical techniques for each of Optical Flow and Hand Tracking (to assess H2). To make

these four comparisons we used nonparametric Wilcoxon Signed Rank Tests (paired version, 1-sided).

When comparing the Predictive and Historical techniques, we found a significant difference for Optical Flow (V=1.5, p <0.02). For Hand Tracking, the difference was marginally significant (V=6.0, p <0.1). Results for the comparison of Optical Flow to Hand Tracking were not significant.

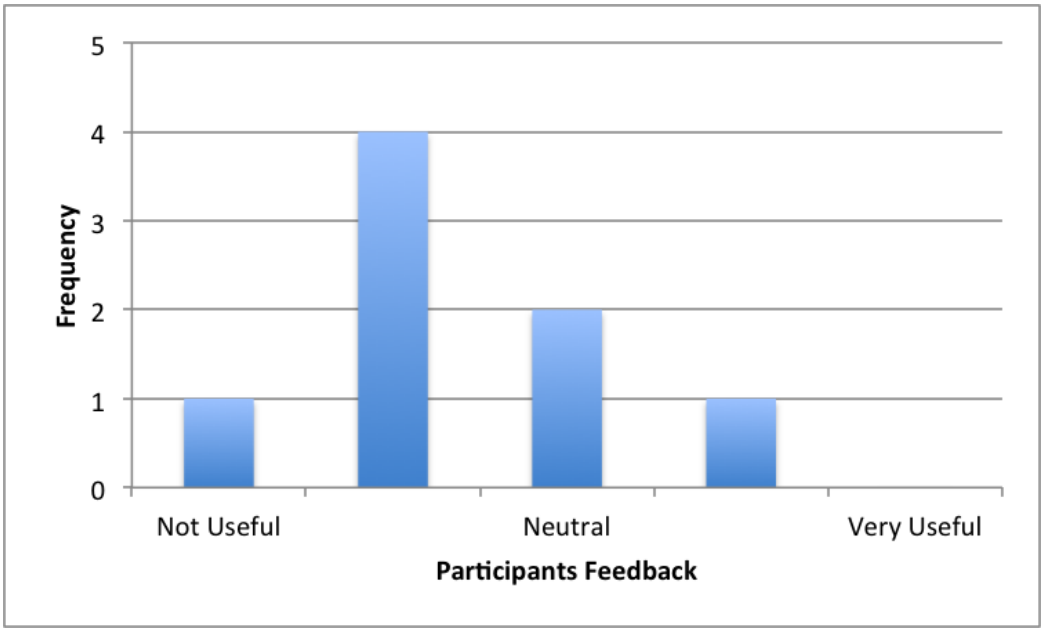

Figure 4.2: Predictive Optical Flow visualization.

Figure 4.2 displays a histogram of the participant's feedback for the predictive optical flow augmentation. As can be seen it is unimodal with the median and mode values between not useful and neutral. Some of the participant's free-form feedback for this augmentation includes:

- "Was not accurate enough." Participant 5.
- "A bit confusing some lack of clarity." Participant 6.
- "Movement not clear." Participant 7.
- "Seeing how much the arm moves is really useful." Participant 4.
- "Did not stay with me." Participant 8.

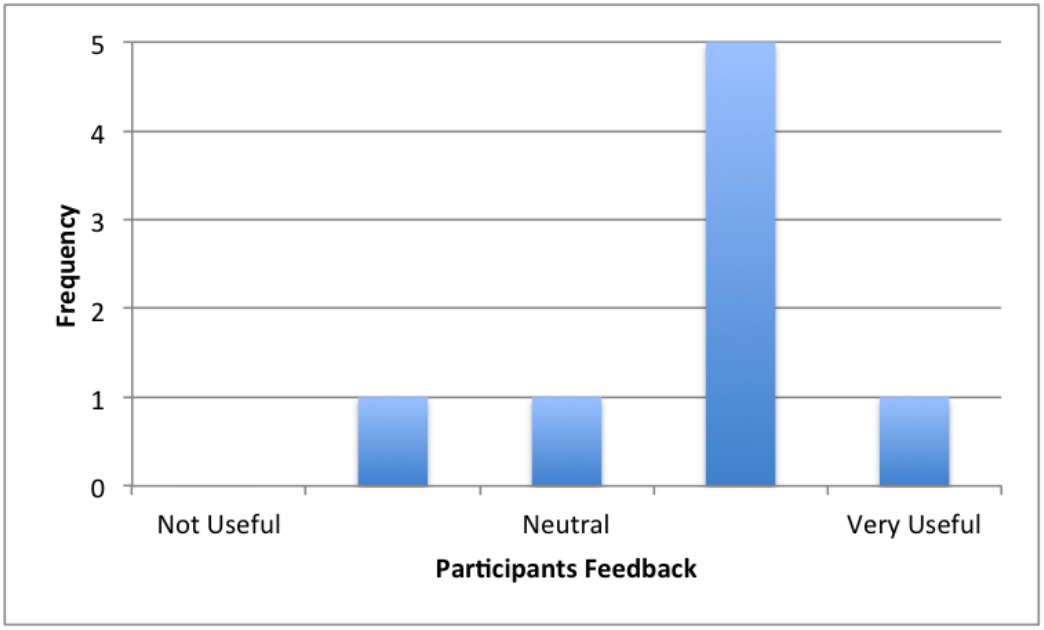

Figure 4.3: Historical Optical Flow Histogram.

Figure 4.3 displays a histogram of the participant's feedback for the historical optical flow augmentation. As can be seen it is unimodal with the median and mode values between neutral and very useful. Only one participant that rated the augmentation as useful responded to the feedback section of the survey. The free-form feedback included:

- "I liked this one the best but not sure why, seemed to appeal to a more abstract way of thinking." Participant 8.
- "If more accurate might be better." Participant 5.
- "Very confusing, too much info." Participant 6.
- "Cleaning up the movement and allowing it to linger longer would be helpful (fade?). I did not like the color red." Participant 7.

Figure 4.4 displays a histogram of the participant's feedback for the predictive hand tracking augmentation. As can be seen there is a large degree of variability in the responses to this visual augmentation. This was the only augmentation that at least one participant chose each value on the Likert scale. The free-form feedback for this augmentation included:

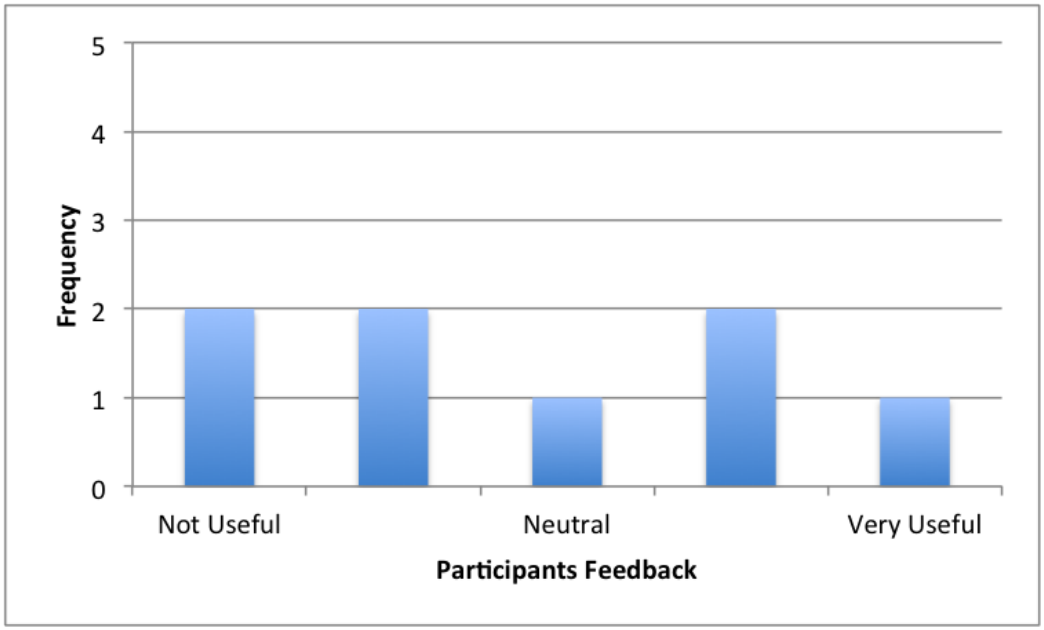

Figure 4.4: Predictive Hand Tracking Histogram.

- "Showed the stillness of the keyboard player. The efficiency of the movement. Small movement of the head and eyes. Stillness of the resting hand." Participant 8.
- "Hard to see movement and hand motions vary greatly by each technical approach." Participant 4.
- "Movement not clear." Participant 7.

Figure 4.5 displays a histogram of the participant's feedback for the historical hand tracking augmentation. As can be seen it is unimodal with the median and mode values centralized between neutral and very useful. The participants that rated the augmentation the highest did not respond when asked to comment on the augmentation. The free-form feedback included:

- "It was difficult to only judge the tracking as watching Helmut Brauss was extremely helpful in all videos." Participant 5.
- "Clean up the movement and allow it to linger longer would be helpful." Participant 7.

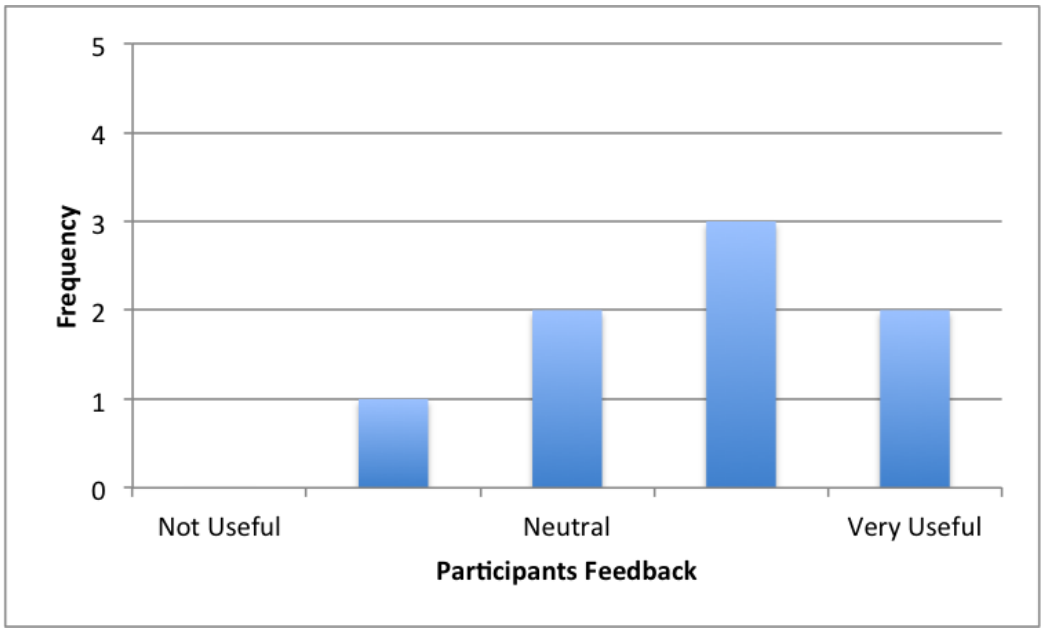

Figure 4.5: Historical Hand Tracking Histogram.

4.3 Discussion

H1 predicted that the hand tracking visual augmentations would be rated as more useful by the students than the optical flow augmentations, because of their simplicity. The experts we interacted with felt that the optical flow visual augmentations drew too many feature vectors over the video and as a result would be confusing to students. However, the results of the user study could not adequately assess this hypothesis. No significant difference was found between the participants' preference for the hand tracking and optical flow visual augmentations. While there may be a difference between these augmentations, we were not able to observe it with our small sample size.

H2 predicted that participants would prefer the historical to the predictive visual augmentations. The experts we interacted with were concerned that the information presented to the students in the predictive augmentations was only briefly visible. The historical information stayed visible for a longer period of time and was therefore thought to be easier to interpret. The results of the user study were consistent with the experts' evaluation of the visual augmentations. The students surveyed consistently preferred the historical to the predictive visual augmentations (significant difference for optical flow and marginally significant for hand tracking). The pre-

dictive augmentations are best viewed in slow motion or controlled frame by frame by the student. Unfortunately when the videos are viewed in this fashion it is not possible to hear the performer playing the Étude. This could be a contributing factor to the difference found between the augmentations.

Overall, our study demonstrates that students found the augmentations useful, particularly the historical ones. Nonetheless, this study is very preliminary and subject to many caveats. The small sample size of the study limits its generalizability, and it is certainly possible that a wider population may have differing opinions on the augmentation. Nonetheless, we are fortunate to have received feedback from this many participants given the specialized and experienced nature of the population. Also the decision to have students view the videos and complete the questionnaire at home reduced experimental control. For example, it is possible that some videos may have been viewed for a longer time than others, and likely that different participants devoted different amounts of time and effort to the evaluation. Nonetheless, this approach was necessary to avoid taking valuable classroom time for the study.

When designing the user study, we originally intended to include a control group to determine the usefulness of the visual augmentations. The participants would be broken into two groups: one that had access to the visually augmented videos and one that did not. The students would then proceed through the class and the usefulness of the augmentations would be determined by the instructor. The authors hoped that the instructor could evaluate the students and determine if the students who had access to the visual augmentations were more successful then the students who did not. There are a number of reasons the authors decided not to pursue this type of study. First, the instructor informed the authors that it would be difficult to correlate the students' performance with viewing the visual augmentations. Many of the students had different levels of Chopin experience and no previous contact with the instructor so it would be difficult to asses if the visual augmentations improved their ability to learn the Études. Secondly, it would be difficult to guarantee that students who had access to the visual augmentations did not share them with students who did not have access to them. Finally, because of the limited number of participants available, there were too few participants to create a control group.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this work, we introduced four visual augmentations to videos designed to help students learn Chopin's Études. Students found some of the visual augmentations to be more useful than other visual augmentations. To the best of our knowledge no other study has been conducted in this area. The visual augmentation of Chopin's Études was used for this study but the augmentations can be generalized to many applications.

Students in our study preferred the historical augmentations to the predictive augmentations, possibly because the predictive augmentations are best viewed in slow motion or frame-by-frame, when the audio cannot be heard. It is important to note that the historical visualizations stayed visible for a longer period of time and that could be a strong factor in determining the users preference. We found no difference between the optical flow and the hand tracking visual augmentations. This fact is contrary to the expert's initial intuition that students would find the hand tracking visual augmentations more useful than the optical flow based visual augmentations; however, such a difference might be found with a larger sample size.

5.2 Future Work

In this work markerless techniques were used to create the visual augmentations of Helmut Brauss performing Chopin's Études. This choice was made because the videos existed prior to the research being conducted. More accurate information could be

achieved by using a marker based motion capture system. The Études would have to be re-recorded but the quality of the information would be significantly improved. Important body points like the head, shoulders, elbows could be explicitly marked. Most modern motion detection systems operate at 120 progressive frames per second this would be a significant improvement to the 30 interlaced frames per second used in this project.

Having the predictive visualizations visible on the screen for a longer period of time could provide interesting results in a future study. This could be achieved by simply inverting the historical visualization. That is, instead of drawing where the motion "was" draw where the motion is "going". The motion data is available for the entire video sequence so it is possible to show the user where the motion will be in the future.

The data used in this project was displayed visually to the user, but after a calibration process it could be stored and compared to future sessions. It would be extremely interesting to film a student in the early stages of learning Chopin's Études and then at regular intervals as they advance to become expert players. It would be possible to quantitatively track the motion differences in that particular student as they progressed to a Chopin Étude expert.

We used the visual augmentations to teach students Chopin's Études, but the visual augmentations are not limited to this specific activity. The only precondition is that the camera remains fixed throughout the video sequence for optical flow based visualizations. With this precondition in mind the visual augmentations can be used to teach any motion based activity. The most logical activity extension would be into the sports world. The visual augmentations could be used for sports like golf or baseball as long as a sufficient frame rate is used related to the activity being performed.

The two dimensional histogram has a variety of potential uses because it contains a compact representation of the motions necessary to perform each Étude. Normalizing this data and correlating it with MIDI information similar to Oka and Hadjakos [31, 17, 16] would be an interesting research project. Machine learning algorithms could be used to cluster motions between multiple subjects and a framework for the general motions necessary to perform Chopin's Études could be created.

Appendix A

Additional Information

A.1 Survey Recruitment Script

Michelle Mares will say, "Welcome to class. As you all know the Computer Science department provided the augmented videos used in this class. They are interested in getting some feedback to determine if the videos helped you learn Chopins Etudes. The information they gather will be used to evaluate the usefulness of the augmented videos and will be published in David Kerrs Masters thesis and in a scientific journal. Your participation is not mandatory and we have decided to conduct an implied consent questionnaire. I will hand out an implied consent form and the questionnaire. If you decide to complete the questionnaire you will have given consent to use the information for research purposes. If you decide not to participate in the project all you need to do is not return the questionnaire. Your participation will not affect your mark in this class and it will have not affect on you class standing, your relationship with Michelle Mares and your relationship or standing with the Victoria Summer Piano Academy or UVic. We have structured the questionnaire so that they are anonymous and after you have submitted the form there is no way to withdraw. The best way to submit the form is to return it to me during class but you can also email it to dkerr@uvic.ca or mail it to the Computer Science department care of David Kerr. Are there any questions?"

A.2 Survey Questionnaire

1. What is your year of birth?
2. How many years of experience do you have playing piano?
3. How many months have you been playing Chopins etudes?
4. How many hours a week do you practice piano on average?
5. On a scale of one to five, was the predicative flow augmentation useful when learning Chopins etudes? (Many green vectors)

   (Not Useful) 1 2 3 4 5 (Very Useful)
6. Please provide any feedback about the predicative flow augmentation:
7. On a scale of one to five, was the historical flow augmentation useful when learning Chopins etudes? (Many red vectors)

   (Not Useful) 1 2 3 4 5 (Very Useful)
8. Please provide any feedback about the historical flow augmentation:
9. On a scale of one to five, was the predicative hand tracking augmentation useful when learning Chopins etudes? (Green hand vectors)

   (Not Useful) 1 2 3 4 5 (Very Useful)
10. Please provide any feedback about the predicative hand tracking augmentation:
11. On a scale of one to five, was the historical hand tracking augmentation useful when learning Chopins etudes? (Red hand vectors)

    (Not Useful) 1 2 3 4 5 (Very Useful)
12. Please provide any feedback about the historical hand tracking augmentation:

Bibliography

[1] C. Barron and I.A. Kakadiaris. Estimating anthropometry and pose from a single image. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 669–676 vol.1, 2000.

[2] Carlos Barron and IoannisA. Kakadiaris. On the improvement of anthropometry and pose estimation from a single uncalibrated image. *Machine Vision and Applications*, 14(4):229–236, 2003.

[3] A. Bissacco and S. Soatto. Classifying human dynamics without contact forces. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1678–1685, 2006.

[4] Oren Boiman and Michal Irani. Detecting irregularities in images and in video. *Int. J. Comput. Vision*, 74(1):17–31, August 2007.

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] M. Brand. Shadow puppetry. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1237–1244 vol.2, 1999.

[7] Helmut Brauss. *The Pianist's Breviary.* University of Alberta, 2013.

[8] Christoph Bregler, Jitendra Malik, and Katherine Pullen. Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, 56(3):179–194, 2004.

[9] Antonio Camurri, Barbara Mazzarino, and Gualtiero Volpe. Analysis of expressive gesture: The eyesweb expressive gesture processing library. In *In Gesture-based Communication in Human-Computer Interaction, LNAI 2915*, pages 460–467. Springer Verlag, 2004.

[10] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, July 2003.

[11] Ginevra Castellano, Marcello Mortillaro, Antonio Camurri, Gualtiero Volpe, and Klaus Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, 2008.

[12] Amit K Roy Chowdhury and R. Chellappa. A factorization approach for activity recognition. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, volume 4, pages 41–41, 2003.

[13] Howlung Eng, K.-A. Toh, A.H. Kam, J. Wang, and Wei-Yun Yau. An automatic drowning detection surveillance system for challenging outdoor pool environments. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 532–539 vol.1, 2003.

[14] Werner Goebl and Caroline Palmer. Temporal control and hand movement efficiency in skilled music performance. *PloS one*, 8(1):e50901, 2013.

[15] D.O. Gorodnichy and A. Yogeswaran. Detection and tracking of pianist hands and fingers. In *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, pages 63–63, 2006.

[16] Aristotelis Hadjakos. Pianist motion capture with the kinect depth camera. In *Proceedings of the 9th Sound and Music Computing Conference (SMC 2012)*, pages 303–310, Copenhagen, Denmark, 2012.

[17] Aristotelis Hadjakos, François Lefebvre-Albaret, and IRIT Toulouse. Three methods for pianist hand assignment. In *6th Sound and Music Computing Conference*, pages 321–326, 2009.

[18] N.R. Howe. Silhouette lookup for automatic pose tracking. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pages 15–22, 2004.

[19] Auke Jan Ijspeert, Jun Nakanishi, and Stefan Schaal. Movement imitation with nonlinear dynamical systems in humanoid robots. In *In IEEE International Conference on Robotics and Automation (ICRA2002*, pages 1398–1403, 2002.

[20] O.C. Jenkins and M.J. Mataric. Deriving action and behavior primitives from human motion data. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, volume 3, pages 2551–2556 vol.3, 2002.

[21] R. Kehl, M. Bray, and L. Van Gool. Full body tracking from multiple views using stochastic sampling. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 129–136 vol. 2, 2005.

[22] J. Kovac, P. Peer, and F. Solina. 2d versus 3d colour space face detection. In *Video/Image Processing and Multimedia Communications, 2003. 4th EURASIP Conference focused on*, volume 2, pages 449–454 vol.2, 2003.

[23] N. Krahnstoever and R. Sharma. Articulated models from video. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–894–I–901 Vol.1, 2004.

[24] L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: establishing a common coordinate frame. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):758–767, 2000.

[25] Mun Wai Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–334–II–341 Vol.2, 2004.

[26] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[27] Antonio S. Micilotta, Eng Jon, and Ong Richard Bowden. Detection and tracking of humans by probabilistic body part assembly. In *Proc. of British Machine Vision Conference*, pages 429–438, 2005.

[28] Thomas Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.

[29] Thomas B. Moeslund, Adrian Hilton, and Volker Krger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(23):90 – 126, 2006. ¡ce:title¿Special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour¡/ce:title¿.

[30] R. Navaratnam, A. Thayananthan, P. Torr, and R. Cipolla. Hierarchical part-based human body pose estimation. In *Proc. BMVC*, pages 47.1–47.10, 2005. doi:10.5244/C.19.47.

[31] A. Oka and M. Hashimoto. Marker-less piano fingering recognition using sequential depth images. In *Frontiers of Computer Vision, (FCV), 2013 19th Korea-Japan Joint Workshop on*, pages 1–4, 2013.

[32] Eng-Jon Ong and Adrian Hilton. Learnt inverse kinematics for animation synthesis. *Graph. Models*, 68(5):472–483, September 2006.

[33] V. Parameswaran and R. Chellappa. View independent human body pose estimation from a single perspective image. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–16–II–22 Vol.2, 2004.

[34] R. Plankers and P. Fua. Articulated soft objects for multiview shape and motion capture. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1182–1187, 2003.

[35] Deva Ramanan, D.A. Forsyth, and Andrew Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, pages I: 271–278, 2005.

[36] Y. Ricquebourg and P. Bouthemy. Real-time tracking of moving persons by exploiting spatio-temporal image slices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):797–808, 2000.

[37] Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9):661–670, 2001.

[38] Timothy J. Roberts, Stephen J. McKenna, and Ian W. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In *ECCV*, pages 291–303, 2004.

[39] N. Robertson and I. Reid. Behaviour understanding in video: a combined method. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 808–815 Vol. 1, 2005.

[40] Myung-Cheol Roh, Bill Christmas, Joseph Kittler, and Seong-Whan Lee. Robust player gesture spotting and recognition in low-resolution sports video. In Ale Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3954 of *Lecture Notes in Computer Science*, pages 347–358. Springer Berlin Heidelberg, 2006.

[41] Remi Ronfard, Cordelia Schmid, and Bill Triggs. Learning to parse pictures of people. In *In European Conference on Computer Vision*, pages 700–714, 2002.

[42] Rómer Rosales and Stan Sclaroff. Learning and synthesizing human body motion and posture. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, FG '00, pages 506–, Washington, DC, USA, 2000. IEEE Computer Society.

[43] Naotaka Sakai, Michael C. Liu, Fong-Chin Su, Allen T. Bishop, and Kai-Nan An. Hand span and digital motion on the keyboard: Concerns of overuse syndrome in musicians. *The Journal of Hand Surgery*, 31(5):830 – 835, 2006.

[44] Cristian Sminchisescu and Bill Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22:2003, 2003.

[45] Chris Stauffer and W. E L Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, pages –252 Vol. 2, 1999.

[46] N. Vaswani, A.R. Chowdhury, and R. Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–633–40 vol.2, 2003.

[47] S. Wachter and H.H. Nagel. Tracking of persons in monocular image sequences. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 2–9, 1997.

[48] Brenda Wristen. Sight-reading versus repertoire performance on the piano: A case study using high-speed motion analysis. *Medical Problems of Performing Artists*, 21(1):10–16, 2006.