A Shape-Based Approach to Computer Vision Musical Performance Systems

Jean-Marc Pelletier
Institute of Advanced Media Arts & Sciences
3-95, Ryoke-cho, Ogaki City, Gifu, Japan
jovan02@iamas.ac.jp

ABSTRACT

In this paper, I will describe a computer vision-based musical performance system that uses morphological assessments to provide control data. Using shape analysis allows the system to provide qualitative descriptors of the scene being captured while ensuring its use in a wide variety of different settings. This system was implemented under Max/MSP/Jitter, augmented with a number of external objects. (1)

Keywords

Computer vision, image analysis, morphology, MaxMSP, musical control

1.BACKGROUND

My goal, in this project, was the realization of a musical performance system that would be based on qualitative measurements while being open-ended enough to be used in a wide variety of different settings. To this end, the system proposed uses real-time image analysis as it allows a great level of portability and permits more elaborate data analysis techniques. Vision-based systems like STEIM's BigEye(4) and David Rokeby's VNS(2) offer a good level of open-endedness. However, I felt that their approaches, based on movement and location, could be expanded by higher-level analysis of the properties of the control objects themselves.

2.SYSTEM OVERVIEW

Images captured with any digital camera, or frame grabber are first sent to a filtering and segmentation module. This section of the program selects objects based on color or intensity and applies some amount of noise reduction. Connected components are identified; each individual component, or object, is used to control its own voice, or channel, allowing for multiphonic control. Each connected component is then sent to shape analysis modules. Various qualitative descriptors are computed and then sent out using MIDI, or received by other Max/MSP modules.

3.IMAGE AQUISITION, FILTERING AND SEGMENTATION

Since the image analysis software is built using Jitter objects, image input can be achieved using any camera or capture device supported by QuickTime. Latency, however, is always problematic. PCI capture cards typically offer the best performance, but latencies under 100 ms are still difficult to achieve in the current state of the system.

The filter module proposes two approaches for the identification of objects of interests. The first is a simple intensity boundary method. The image is converted to grayscale and pixels whose intensity fall within user-defined bounds are marked as positive. The second technique uses hue to identify positive areas. The image is converted from RGB to HSL format and only the hue channel is used. Again, the user specifies bounds within which pixels are labeled as positive.

Since the resulting image is often somewhat noisy, and irregular edges can cause the subsequent calculations to be imprecise, the binary image computed above is sent through a noise filter that removes smaller specks and smooths the edges.

Following these primary image processing steps, individual components are identified and separated. Up to eight connected components are given individual labels provided they are larger than a certain value. Again, the number of components identified and the size threshold are user-defined variables. In the case where the number of components that meet the size criterion is larger than the maximum number of objects sought, larger components are given priority.

4.SHAPE ANALYSIS

Once individual objects have been identified and given separate values, shape analysis is performed on them using moment-based algorithms. In moment-based shape analysis, the image is seen as a two-dimensional random distribution. Moments (M_{pq}) are calculated using the formula:(3)

$$M_{pq} = \Sigma\Sigma\, x^p y^q f(x,y)$$

Since we are using binary images, *f(x,y)* equals 1 for pixels that are part of the shape and 0 if they are outside. p and q are positive integers. The order of moments is given by the sum of p and q. For shape analysis computing moments of second and third order is sufficient, in other words:

$$M_{20}, M_{02}, M_{11}, M_{21}, M_{12}, M_{30}, M_{03}$$

These descriptors are normalized relative to size and position to obtain *central normalized moments*. This ensures that measurements are invariant with regard to translation along any axis (movement along the *x*-axis correlates to changes in the visual object's size).

These moments provide, as-is, qualitative description of the objects' shapes. For instance, M_{11} is a measure of skew, while M_{30} measures asymmetry along the *y*-axis. However, they are not invariant with regard to rotation, which is problematic for the kind of open-ended system proposed herein. A number of higher-level descriptors are thus computed from those basic *central normalized moments*.

These descriptors have been chosen for their correlation with obvious visual features of the shapes being used. They include:

Area: This is simply a measurement of the object's size, in pixels. Changes in area either correspond to movement along the z-axis (relative to the camera) or growing and shrinking transformations.

Circularity: This is an assessment of how compact a shape is. Shapes with high circularity, like circles and squares, are very dense, or closed. As circularity decreases, shapes tend to form extensions and spokes; the outline becomes more intricate and the shape can be seen as opening.

Elongation: This descriptor relates to how concentrated along its main axis a shape is. Rounder shapes will yield low elongation values. Elongation increases the more shapes tend towards thin, straight lines.

Orientation: This is a calculation of a shape's main axis' angle. For shapes that display low elongation, this descriptor is somewhat less relevant, but for thinner shapes, orientation describes rotations quite accurately.

The values computed are then mapped to integer values using user-defined tables. Each individual component's data is output on a separate channel, sent either to a MIDI device or a Max “receive” object. Polyphonic control can thus be achieved but, in the current implementation, with a caveat: the component labeling scheme is not temporally consistent. A component is not guaranteed to always have the same label, and hence, its shape data may not always be sent out the same channel.

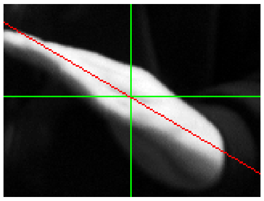
Furthermore, the system can be trained to recognize up to seven different shapes. Program change messages are sent out to the appropriate channel when a particular shape has been positively identified. Even though the robustness of shape recognition is greatly dependant on the quality of the training set, near 100% positive identification can be achieved in ideal circumstances.

5.IMPLEMENTATIONS

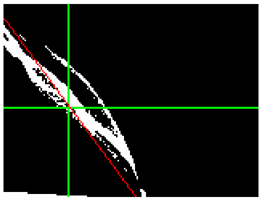
Versions of this system have already been implemented in a few very different settings. Notably, my piece *Kimi ni Yuehen Koto* uses hand postures to control a musical performance. Shape recognition is used to trigger spoken syllables. Musical production algorithms are changed when complete words are formed. Individual timbre parameters can be controlled by the performer only slightly moving his fingers.

The same system was also used in the very different *Kemuri-mai*. In this piece, the smoke rising from a burning stick of incense was used to control the musical score. Since in this situation, shape recognition is of little meaning, only real-time parameters were used. As the smoke sways, expands, contracts, is disturbed by air turbulence, these changes are mirrored in the soundtrack.

That the system could be used in such different pieces with no modification at all shows that the goals of portability and open-endedness have been achieved. Furthermore, in both pieces, shape analysis allowed the sounds to mirror actual qualities of the visual objects, thus avoiding overly arbitrary mappings.



In this scene, an outstretched hand results in high elongation values. Orientation (shown by the red line) can be easily controlled. Slightly bending the fingers results in small changes in circularity that can be used for a convincing musical effect.



Here, in this frame from *Kemuri-mai,* the smoke displays a rather average form. It could, from then either expand, or contract to a thinner column, which would be reflected in elongation values. Orientation changes can be used not only to assess changes in angle, but also in curvature.

6.FUTURE WORK

Technically, this system can be improved by providing more elaborate or efficient methods of image filtering and segmentation. Especially useful would be some sort of temporal consistence to the labeling of connected components. However, the most important aspect of future research should be the development of a solid conceptual and aesthetic framework for the pairing of shape changes to sound. Implementation of new shape descriptors should derive from the insights drawn from a better understanding of audio-visual relationships.

7.REFERENCES

[1] Pelletier, J.-M. *cv.ji* Max/MSP/Jitter externals available on the WWW at http://www.iamas.ac.jp/~jovan02/cv/

[2] Rokeby, D. *Very Nervous System* WWW page at http://homepage.mac.com/davidrokeby/vns.html

[3] Seul, M., O'Gorman, L., Sammon, M.J. *Practical Algorithms for Image Analysis: Description, Examples and Code.* Cambridge University Press, 2000

[4] STEIM *BigEye* Software available on the WWW at http://www.steim.org/steim/bigeye.html.