Sound and Sight:
Composition for Image Analysis Systems

Thesis by
Jean-Marc Pelletier

Submited to:
Institute of Advanced Media Arts & Sciences
2004-1-29

Introduction

With the continual inflation in computer performance, the possibilities for real-time control of musical performance are rapidly expanding. These new possibilities allow for the development and growth of not only new devices but of completely new control paradigms that go beyond the traditional concepts of instrument and controller. However, with new possibilities also come new responsibilities. Creators must understand the workings and effects of their new tools not only on their basic, technical dimensions but most importantly at the aesthetic level.

Specialist and generalist systems

Musical performance systems can be categorised according to various criteria, however, they generally fall within two approaches. The first is the *specialist* approach. This category of performance system includes the most traditionalist, instrument-like devices, like Tod Machover's *hyperinstruments*.[1] Systems can be *specialist* either relative to the source of control information, or to the parameters being controlled. This type of approach typically allows for great sensing and control accuracy, as sensors are usually created specifically for a narrow range of phenomena, as well as specific control parameters.

The pendant to *specialist* approaches are *generalist* systems. These types of system should allow for the use of a wide range of controller data. They typically offer greater creative possibilities, although in many cases at some cost in accuracy of measurement. An early example of such approach is the famous *Variations V* by Merce Cunningham and John Cage.[2] The system, designed by David Tudor and Gordon Mumma triggered sounds when dancers approached a number of electronic sensors placed about the stage. Even though the system was created for a dance performance it is *generalist* because of its open-ended design; it will react to a great number of different of behaviours, unlike *hyperinstruments*, that are built to monitor a very narrow range of performer actions. One particularly suitable technological approach to *generalist* systems is through the use of computer vision, or real-time image analysis.

Composition and image analysis

There have already been a few attempts at using computer vision systems for the control of musical performance. However, only recently have computer capabilities allowed for the practical implementation of the more advanced analysis techniques for musical purposes. As such, there is still a relative dearth of critical thinking about the aesthetic implications of these technologies.

In this thesis, I propose a few simple aesthetic concepts to assist the act of composing for real-time image analysis systems. Since these systems extract useable information from visual data, I believe that composing for them should be seen as the act of pairing auditory and visual information. In order to understand how these audio-visual pairs can function, I call upon elements of Rudolph Laban's theory of movement, as well as part of Pierre Schaeffer's views on the different levels of listening. By combining these two points of view – one for the interpretation of visual information, the other for the comprehension of auditory elements – I hope to provide a simple framework for the analysis and use of visual control of sound.

(1) Machover, Tod *Hyperinstruments* http://brainop.media.mit.edu/Archive/Hyperinstruments/index.html
(2) Nyman, Michael *Experimental Music: Cage and Beyond* (p. 97)

In the first section, I introduce the very basics of Laban's theory of movement. I will propose a particular interpretation of, and modifications to this theory that will allow its use in the most generalist of contexts. Essentially, I wish to move musical control paradigms away from human body-specific paradigms, to allow for the use of any external phenomenon as source of control data. In the second section, I briefly introduce the elements of Schaeffer's theory that I use as the base for my framework. In the following, central section, I discuss the aesthetics of audio-visual pairing using the basic concepts derived from the theories introduced previously. I then move on to an explanation of some of the shape analysis techniques available. Shape analysis provides us with data that can be readily and easily adapted to implement the approaches outlined in previous sections. I then overview some of the existing image analysis software for artists and outline some of their strengths and weaknesses relative to the goals of this thesis. Finally, I propose an example of how our framework has been implemented. My piece *Kemuri-mai* is a musical performance that uses smoke to control sound synthesis. Such a piece would be difficult to analyse or conceive using the traditional, human-oriented paradigms, both on technical and aesthetic levels. However, using our generalist approach I can easily outline some of its outstanding features, as well as detail the methodology used in its creation.

Laban Movement Analysis

Fundamentals

Laban Movement Analysis (LMA) is a framework for the description, notation and interpretation of human motion. It originated in the first half of the 20th century from the work of dancer, theorist and choreographer Rudolph Laban (1879-1958) and was expanded on by his colleagues and students. LMA is broadly used in the fields of dance and acting but has also been used for such diverse purposes as therapy and ergonomics.[1]

There have been a number of attempts at integrating elements of Laban Movement Analysis in computer vision applications. They include tools for high-level analysis of human gestures, for instance, Antonio Camurri's work with the *EyesWeb* software.[2] Another project, *EMOTE*, uses LMA to both analyse and synthesise gestures with the goal of adding a higher level of expressivity to computer-generated characters.[3]

Even though Laban Movement Analysis was conceived for the study of human motion, some of its concepts can be applied to the more general approach to image analysis that is proposed herein. Owing in part to the nature of Laban Movement Analysis, many of its tools can be equally well used for the description and interpretation of human and non-human movement alike. It should be noted, however, that the goal here is only a partial adaptation of the framework, not a complete generalisation of LMA.

At its most basic level Laban Movement Analysis is composed of five components: *Body, Space, Effort, Shape* and *Relationship*. These components can be seen as perspectives from which a movement can be studied. LMA proposes both the study, in depth, of each of these components and the study of how they integrate with each other. Some researchers, however, chose to mainly focus on one or two components. This is the case, for instance, of Irmgard Bartenieff's exploration of the Body component.[4]

Components

Body deals with the anatomy of movement, in other words a description of movement relative to the various body parts. *Body* analysis concerns where a movement originates and how it propagates in the body of the mover. For a generalist approach, this can prove problematic. For some objects, such as those that contain no moving parts, it might not be relevant at all. For other objects and phenomena, a custom language, or a set of specific criteria might be developed to properly understand their motion from a *Body* perspective.

Space is the description of the *kinesphere* the moving object occupies; the realm of possible and likely locations of a moving object. As an object moves in an environment, it creates in the viewer a certain expectation as to where it ought to be able to move to – this is what is referred to as the kinesphere. Furthermore, as objects move about, they have the tendency to draw out certain forms in space. A good example of this would be people drawing Chinese ideograms in thin air. The motion creates a virtual static image, from the viewer's perspective.

Effort concerns the quality of movement and the subjective appraisal of the use of energy by the object in motion. Of the five motion

(1) Zhao, Liwei *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters* (p. 38)
(2) Camurri, Antonio et al. *EyesWeb - Towards Gesture and Affect Recognition in Dance/Music Interactive Systems*
(3) Chi, Diane et al. *The EMOTE Model for Effort and Shape*
(4) Bartenieff, Irmgard *Body Movement: Coping with the Environment*

components, it is the most qualitative in nature. A number of *Effort* factors are outlined and movements are said to indulge or resist extreme tendencies of these factors.

Shape looks at the morphological transformation of the object. It is a more general framework than the *Body* perspective and focuses on the changing relationship of an object with its own form.

Lastly, *Relationship* analyses the interaction of the object with others and with its environment.

Of these five components, *Effort* and *Shape* are the best suited for audio-visual analysis and pairing. Liwei Zhao in the EMOTE project also focused on these two perspectives[1] as they provide the most general and useable analysis data for the purpose of interface design.

Here follows a more detailed description of the *Effort* and *Shape* components of Laban Movement Analysis.

Effort

Effort is defined by four motion factors: *Space, Weight, Time* and *Flow*. Movements can be graded according to the extreme states of these analysis criteria.

Space

Space is the assessment of the trajectory's qualities. A movement is said to be *indirect* if its trajectory is multi-focal. In subjective terms this would correspond to wandering, or tangled motions. Examples of *indirect* movement include a butterfly's flight and the swaying of wind chimes. *Direct* movements, on the other hand, have a single focus and strong directionality – like a train or the arc of a thrown baseball. The *Space Effort factor* should not be confused with the *Space* component. The *Effort* factor of *Space* is a description of the spatial quality of movement whereas the *Space* component is a physical description of the space an object occupies.

Weight

Weight can be thought of as the perceived amount of energy a moving object possesses. *Light* movements are those that display a propensity to go against the pull of gravity and low inertia. *Strong* movements feature high kinetic inertia, and appear very energetic. *Weight* seeks to quantify the perceptual difference between a punch and a gentle outstretching of the hand.

Time

To borrow a sonic term, *Time* is the study of a movement's envelope. Movements can be thought of as being either *sustained* or *sudden*. *Sustained* movements generally display smooth attacks with prolonged sustain portions. In some cases, *sustained* motions might not even have noticeable attack or release components, such as with the seemingly constant flow of traffic. *Sudden* movements, on the other hand, have sharp attacks and releases, such as the slamming of a door.

Flow

Flow is a measure of the amount of control that is being exerted on an object in motion. A *free* motion is one that displays a lack of control. In other words, it is outside of the realm of influence of its originator. This would be the case of a thrown object, for instance; once the movement initiated, it cannot be altered by the thrower. The opposite are *bound* movements, where the object continuously remains within the sphere of influence of whatever sets it in motion.

(1) Zhao, Liwei *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters* and Chi, Diane et al. *The EMOTE Model for Effort and Shape*

Shape

The *Shape* component of motion is divided in three different qualities that describe the ways in which an object may transform: *shape flow, directional movement* and *shaping.*

Shape flow

Shape flow is primarily the study of a shape outline's changes relative to its perceived centre. In proper Laban Movement Analysis, shape flow is approached from two perspectives: that of the torso and that of the limbs. *Shape flow* can, however, be easily generalized. *Growing* and *shrinking* are shape transformations that correspond to broad contractions and expansions of the object's outline. In two-dimensional computer vision this can prove problematic as growing and shrinking shape flows can easily be confused with movement along the *z*-axis. A shape can also be seen as *opening* or *closing*. In a human body, this corresponds to movement of the limbs but the terms can also be generalized as complexification and simplification of the shape outline. As a shape *opens*, it forms spokes and extensions which has the effect of lengthening its perimeter. The contrary motion, as a shape *closes,* is that towards a denser shape, or as will be defined in a later section, augmenting its *circularity*.

Directional movement

Directional movement is defined as a transforming object's tendency to bridge two points in space. When someone outstretches an arm, not only is that person opening, she is moving towards something, or in a given direction. That movement may be seen as *spokelike*, or bridging two points in a direct fashion. An *arclike* movement is one that draws a more indirect trajectory.

Shaping

Shaping is the tendency of an object to adapt to its environment shape-wise. More commonly, an object can be said to "take the shape of something." This can occur through direct interference of other objects, such as someone trying to fit in a narrow space, or two objects deforming as they hit each other. It can also occur independently but in such a way as to suggest the presence of another object or force. This is the effect that mimes achieve when they almost manage to convince their audience of the reality of invisible objects.

Pierre Schaeffer's Levels of Listening

Pierre Schaeffer (1910-1995) is regarded as the originator of the *Musique Concrète* movement and is the author of the seminal *Traité des objets musicaux* (1966). His major contribution can be said to have been the proposition of a general system for the description and analysis of sound. Schaeffer's approach is phenomenological in nature; he states that sound is not the signal but its perception. This perception, while personal, is not entirely subjective. We can talk about the various characteristics of sounds despite those being largely a product of our consciousness. Those characteristics are also not those of the sound wave or the source of the sound but of the sound itself. Schaeffer coins the term "sound object" to distinguish a sound from its source and signal.

At the heart of Schaeffer's theory is his description of the four different ways we can listen to a sound.[1] These four *listenings* are not mutually exclusive and they are not sequential either. They are four different levels at which we can draw information from a given sound object.

In the *Traité des objets musicaux*, Schaeffer uses four French verbs to describe those levels of listening: *ouïr, écouter, entendre* and *comprendre.* This is somewhat problematic as the distinctions between these words cannot be easily translated into English. For this reason, I will substitute for those four verbs the terms: *direct, causal, morphological* and *semantic*. These diverge somewhat from Schaeffer's original intent but they will allow me to use them to describe perceptions that extend beyond sound.

The direct level (*ouïr*)

The direct level of listening is the act of allowing sounds into our minds. This is not an activity initiated by a conscious decision. It is an activity carried on twenty-four hours a day. If a loud sound wakes us in our sleep, it must have entered our mind to do so. In a way, this level of listening is the prerequisite for the remaining three. As such, and since it is largely an unconscious activity, it does not provide as good an analysis tool as the others. As a matter of fact, Michel Chion, a composer, film critic and author, student of Pierre Schaeffer, sometimes refers to the *three* listenings in his writings.[2]

The causal level (*écouter*)

When asked to describe a sound many will start by stating its source. "It's a guitar." "It's the sound of the ocean." "It's someone talking." At this level of listening, sound provides us with clues and information about our environment. This level of listening is said to be objective, in the sense that it carries no value judgement about the sound in question. The information that is being extracted from the sound object refers to an external object or event. When we listen at the causal level, we seek to answer the question: what is making that sound?

The morphological level (*entendre*)

Pierre Schaeffer uses the term *écoute réduite* to refer to the act of paying attention to the qualities of the sound object itself. Those characteristics are not those of the sound source, nor are they those of the physical signal, but of the sound object. Much of Schaeffer's work revolves around establishing a vocabulary and aesthetic framework for the description of sound objects' timbral properties. Culturally, a great part of our sound vocabulary is limited by a very narrow conception of music. We

(1) Schaeffer, Pierre *Traité des objets musicaux* (pp.112-128)
(2) Chion, Michel *L'audio-vision* (pp. 25-31)

either describe sounds in term of their musical parameters, i.e. their pitch or amplitude or their source. This approach falls short of providing an adequate method of describing all sounds. Terms like *mass* and *grain* are thus used to refer to morphological characteristics of sounds.

The semantic level (*comprendre*)

Sound objects can also be used to convey information that is independent of the sound's source or its internal characteristics. In semiotic terms, this would correspond to sound objects being used as *signifiers*. This is of course the level at which we listen to speech. To recognize a sound as being a doorbell is causal listening. However, to think: "there is someone at the door," is semantic. The person's act of standing at the door is not the material cause for the sound and morphological listening excludes any reference to the outside world. There are of course sounds that semantically refer to concepts while sharing some morphological qualities with their referrer. Peirce would call this relationship iconic.[1] The most obvious example of this would be onomatopoeia.

These four types or levels of listening can be extended to sight. The direct level of sight would be the phenomenon of images forming in our minds. This is an activity that continues even when our eyes are closed, though greatly reduced, as some sort of image nevertheless forms. The causal level corresponds to the identification of the image's source. Note that while sound requires some sort of action or transfer of kinetic energy in order to exist (hence the term causal), there is no such requirement for sight. The morphological level of sight is generally much more developed than its sonic counterpart, perhaps owing to having a better vocabulary. We talk about shapes being round or angular, or about objects' colours and textures. Those are all morphological components, in that they describe internal characteristics of the visual object itself. Finally, as with sound, there can be a semantic approach to sight, where images are used as vectors for meaning.

(1) Peirce, Charles Sanders *Collected Papers of Charles Sanders Peirce*

Audio-Visual Relationships

The notion of object

In order to talk about sound and sight relationship, let us review the definition of visual and sound objects. Objects are not their sources, they are not the physical objects that reflect or emit light and they are not the events that set sound waves in motion. Objects are not signals, a visual object cannot be described in terms of light spectrum nor can a sound object be described through measurements of sound waves. Objects require human perception to exist, but they are not entirely in the eye of the beholder. It is possible to talk about objects' characteristics and these characteristics should remain fairly constant from individual to individual even though they are the product of an internal process.

The act of setting sound to sight, in the case of the approached proposed in this thesis, using visual information to control sonic events, is the act of combining visual and sound objects to create multi-dimensional audio-visual objects. Obviously these pairings can occur in endless configurations and possibilities are virtually limitless. There are nevertheless a number of pairing strategies that can be employed, and relationships can be analysed from different perspectives.

Basic parameter mapping strategies

In the study of gestural control of musical parameters, three basic mapping strategies are usually put forward[1]:

One-to-one mapping

In this approach, the simplest, a single analysis parameter is used to control one synthesis parameter. For instance, an object's size might directly affect the sound's amplitude; as the object grows larger, the sound becomes louder.

Divergent mapping

In this type of mapping, one analysis parameter is used to control several synthesis parameters. A visual object's size might be mapped to both amplitude and pitch, for example. The problem with this approach is that it creates a static relationship between more than two performance elements. The expressive possibilities of this relationship are soon exhausted and there is a risk of falling into patterns that are too simple.

Convergent mapping

This strategy utilises the input of several analysis factors to control a single synthesis parameter. Typically, and in the simplest scenario, one parameter will control coarse changes while another one is assigned to finer modulations. More complex and dynamic relationships are also possible.

Empathy

Whenever a visual object and a sound object are perceived simultaneously, they automatically form an audio-visual pair. There will always be some sort of relationship between the two elements of this pair, whether the pairing was intentional or not. One broad concept used to understand these pairings is that of *empathy*.[2]

A relationship is said to be *empathic* when

(1) Rovan, Joseph Butch et al. *Intrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance* (p. 2)
(2) Chion, Michel *L'audio-vision* (p. 11)

changes in one object are mirrored in the other. The three mapping strategies outlined above are all examples of *empathic* pairing. A relationship lacks *empathy* when both objects appear to change independently of each other.

The *empathy* might exist at different interpretation levels. The relationship between a sound and an image might be totally un-*empathic* at the *morphological* level while being very *empathic* at the *semantic* level. The movements and transformations of the visual object would have no bearing whatsoever on the features of the sonic object but the meaning carried by both would be in some sort of agreement.

Empathy also need not be dynamic. That is to say that an empathic relationship might be static, with aspects of constant value mirroring each other.

Levels of interpretation

As has been outlined, Pierre Schaeffer's four levels of listening can be extended to the description of visual objects. This means that those levels can also be used to talk about the various types of relationships formed between visual and sound objects. From a creator's perspective, sight-sound pairings can be decided in any of three possible domains: the *causal*, the *morphological* and the *semantic*. I choose to skip over *direct* listening because it does not allow room for aesthetic decisions.

The conscious mapping of a visual characteristic to a sonic one may be uni-dimensional. In this situation, the mapping occurs in a single perceptive field. Uni-dimensional mapping does not exclude convergent and divergent approaches but it supposes that only semantic elements are used to control semantic aspects at the synthesis stage. A multi-dimensional approach would involve several levels of interpretation. For the sake of simplicity and clarity, we will focus on a uni-dimensional approach, although it should be noted that some amount of "cross-talk" between spheres always occurs.

Causal relationships

Causal mapping plays a large role in filmmaking and is at the very base of foleying work. The objective in empathic causal relationship is to create the illusion that a relationship of cause and effect exists between both sound and sight. In film, this is done through editing, whereas in performance or interactive art, this is accomplished through the use of various sensor technologies. In either case, the goal is to fool the spectator into thinking that a material relationship exists where in fact, there is none.

Causal relationships can follow three basic patterns. In the most common situation, a sonic event is made to appear like the consequence of a visual motion. This approach is logical as it corresponds to the majority of real-life *causal* pairings. The opposite pattern may be chosen, however. The change in the visual object appears to be caused by a sonic event. In practice, this is difficult to implement in a completely automated setting as it usually presupposes that the sonic event is either simultaneous with or precedes the visual event, which is technologically impossible when using the visual field as a source of data. It is an approach, however, that can certainly be weaved into a live piece by either having a human performer or by using a sort of feedback system where not only does the image affect the sound but sonic information is used to control visual content. The last pattern corresponds to the situation where both sight and sound are made to seem to react to a third, outside force.

Achieving this effect is largely a question of context and environment and is not technically much different from the first pattern.

The greatest problem faced when trying to establish a causal relationship is that of synchronisations. In live settings performers have to deal with the latency inherent in the hardware and software systems used. Often, the minimum delays are so large as to completely jeopardise any attempt at causal mapping. The tolerance to "soft-synchronization" is in great part dependent on the nature of the sound and visual objects. Percussive sounds and movements, i.e. those that have very sudden onsets are particularly sensitive to this problem. However, if system latency is properly taken into account and both sonic and visual content is properly chosen in consequence, convincing *causal* relationships may be established.

Semantic Relationships

Even when in the case of a very abstract object that is not meant to convey any particular meaning, there is always some form of semantic information carried by an object or a pair of objects. That is because objects can never be totally isolated from their contexts. Even if an object is not meant to *denote* anything it will always carry some sort of *connotation*. A sound might evoke feelings of tension and an image might seem aggressive even though they are completely abstract entities.

Using Ferdinand de Saussure's terminology[1], a pairing of two objects can be said to creates a *syntagmatic* relationship, that is, that the two elements exist side by side. However, each object also exists in the context of a *paradigmatic* structure. For every sight or sound, there are a number of other objects that could have been used instead or that share some characteristic feature. Hence objects not only form relationships with the objects they are paired with, but also with the objects they *could* have been paired with.

Semantic relationships in an audio-visual pair can be *complementary*; each element brings a different element of meaning and these combine to form new meaning. *Contradictory* mapping involves semantic references that negate each other. Pairing city sounds with visuals that refer to a rural setting might be an example of this approach. It differs from the complementary approach because in the former there is greater cohesion between the two elements. *Convergent* relationships occur when both objects carry very similar information and references. The role of each object in this case is to emphasise the content of the other. Finally, a pairing might be deemed to be *arbitrary*. This occurs when one of the two elements lacks a proper denotation and could easily be substituted by another.

In practice, *semantic* relationships are usually created at the stage of the selection of material. However, it is possible to use techniques such as shape recognition in order to have visual signs trigger semantic content in the soundtrack. Great attention must be given to how the audio elements will colour the perception of visual objects and vice-versa. Objects that would have benign connotations by themselves can take on broad new ranges of meaning through the use of proper *complementary* or *contradictory* mapping.

Morphological relationships

In the following section I will try to show a few of the possible relationships at the morphological level between sight and sound. Although the realm of possible morphological parameters is very large for both sight and

(1) Saussure, Ferdinand de *Cours de linguistique générale*

sound, many do not correlate very well, owing to the very different nature of sight and sound. For this reason, I will focus on the generalised Laban Effort model described previously to propose some ideas relative to the pairing of sonic and visual motion.

The nature of sight vs. the nature of sound

One of the greatest differences between sonic and visual objects is their relationship to energy and change. It can be said that change is the natural state of sound. Sound waves are always the result of a mechanical transfer of energy, which presupposes that sound is always the product of movement. This means that sounds are generally, usually, in a state of continuous morphological change. There are a few sounds that do not appear to change over time. However, when occurring naturally, they can be considered exceptions – for instance the sound of a waterfall – or they are the product of human activity, like mechanical hums and electronic drones. Furthermore, these changeless sounds have a tendency to very quickly fade into the background perceptually, as their information potential (causal, morphological and semantic) is exhausted. On the other hand, the electromagnetic transfer of energy at the source of light waves does not require perceptible movement. This means that visual objects can very well exist in a seemingly static state. The extreme example of these opposing natures of sight and sound is that while it is possible to remove all change in a visual scene by taking a photograph, doing the same with sound would not only be impossible, it would be meaningless. On the other hand, it is very easy to strip all spatial information from a sound by playing it back through a monophonic speaker and it will retain its integrity as a sound. The same cannot be done with visual information, however. A digital image compacted into a single pixel looses all of its information. For this reason, it can be said that the fundamental dimension of sound is time, whereas the fundamental dimension of sight is space.

Time

As outlined above, the fundamental nature of sound objects is change and their natural state is movement. "Movement" here is not to be taken in the spatial sense but in the sense that sounds continually evolve along the time axis. There is another important time-related distinction between sight and sound. Sonic movements are rarely reversible. A sound played backwards is a completely different object from its forward counterpart. Visual movement, on the other hand, can much more easily move in reverse order. There are certainly situations where this is awkward; the movement associated with an object falling is irreversible, just like most sounds. Many movements, however, especially those occurring along the x and z-axes are equally plausible going both ways. Where this distinction of *reversible* and *irreversible* motion is important is in the situation where a creator would attempt to map a visual movement and its opposite to forward and backward versions of the same sound. This pairing is clumsy because while the visual gestures will be perceived to be similar, the same cannot be said of the two sonic objects.

When working with audio-visual pairs, the acuity of perception is also problematic. Our ears pick up time differences with much greater precision than our eyes. A similar link exists for hardware; typically, the latency of audio systems is much lower than that of video devices. This means that the use of individual visual cues to trigger audio events quickly hits the limits of human and machine perception.

When talking about the evolution of

sound in time, the term *envelope* is often used to refer to the changing outlines of a sound parameter, usually amplitude. The same term can be used to describe visual gestures. Here, the *envelope* maps the changes in speed the moving object goes through. A simple, *one-to-one* mapping strategy might involve controlling the envelope of a sound object using the motion envelope of a visual object. This might prove interesting but it is important to remember that, naturally, sounds tend to display sharper attacks and smoother decays. Visual motion, on the other hand, can show very sharp decays but fast attacks are much more rare. Note, that I am talking about mapping a morphological aspect of sight to sound and not trying to create a causal link between the two events, although one might result as a bi-product of the operation.

Another approach would take into consideration the quality of motion. LMA's *sustained* and *sudden* aspects of course have equivalents in the realm of sound. In *empathic* pairing, *sustained* movements might be mirrored through the predominance of soft attacks, while a switch to more *sudden* motions might cause a shift to more percussive sound material.

Space

Although we stated above that the primary dimension of sound is time, that is not to say that sound does not carry important spatial information. Our ears are capable of fairly accurate sound localisation.[1] For certain types of sounds, however, this localisation is difficult. By comparison, the space that a visual object occupies is usually well defined. One of the reasons for this has to do with propagation effects. The boundaries of a given visual object might be ill defined, for instance some clouds, or soft shadows, but they will very rarely extend to encompass the entirety of the visual field. Even in the case of a light source that casts shadows throughout a scene, those shadows are perceived as different objects from the light itself. The sky or walls might be considered as examples of visual objects that can occupy a large part of the visual field. However, since no two objects can be seen to occupy the same segment of the visual field at the same time (a phenomenon referred to as occlusion and very problematic in computer vision) a visual object cannot be said to truly propagate throughout a scene. By comparison, sounds regularly occupy the *sonosphere* in totality or in great part. *Diffuse* sound objects are those that lack proper spatial focus. These might include city hum or wind blowing through the leaves in a forest. These sound objects appear to exist in a multitude of locations simultaneously. They are often caused by the accumulation of a multitude of similar micro-objects. In the extreme cases, like for low sine drones, the localization is so difficult that they seem to occupy the totality of the *sonosphere*. *Expanding* sound objects do have a more or less clear spatial focus, a point of origin, but they expand outward from that point. The major cause for this phenomenon is reverberation. As sound waves bounce on an environment's solid surfaces, they come to the listener from a variety of directions. *Expanding* sounds contain important information about the material characteristics of a space. Simply clapping in a room can tell us a lot about the size, shape and materials of that room. Truly *localized* sounds, which appear to be perfectly constrained in a given space, are so rare that when put in an anechoic chamber that absorbs all reverberations, most people will feel a distinct feeling of unreality. That being said, a sound object that appears to have no noticeable spatial expansion and a very definite location may be labelled as *localized*. These terms are specific to sound objects since, for the reasons given above, all visual objects are localized.

(1) Blauert, Jens *The Psychophysics of Human Sound Localization*

Another important concept for the analysis of sonic and visual spaces is that of imaginary space. When watching a two-dimensional visual representation, like a photograph or a projected movie, our mind reconstructs an imaginary three-dimensional space. Even though, in a movie, the medium only records two-dimensional information, the visual objects in that movie properly occupy a three-dimensional space, albeit a virtual one. Imaginary spaces also come into play when combining visual information with monophonic sound. If a proper causal relationship is established between the monophonic sound and the visual event, the sound will be perceived as originating from the same position as the visual motion. This is the illusion that allows for the dialogue in older movies to seem to originate from the speaker's mouth although the loudspeaker is in a completely different position. This effect can also be used to great effect in live settings and dispenses the use of elaborate sound positioning algorithms.

When trying to use LMA's *Space* information as a source of control information, relationships between sight and sound risk falling into the arbitrary. A possible mapping strategy might involve parameters like melodic contour, but the link is tenuous. *Direct* motion may be correlated with predictable sonic evolution, and *indirect* states with more unpredictable timbral changes. Another approach might use a surround loudspeaker system and control the sonic trajectories according to the quality of visual movement taking place.

Beyond control, however, sound can bring a lot to visual motion. While in the example of the monophonic loudspeaker above, visuals can be used to lend a sound object spatial information, the opposite is also possible. Because the temporal acuity of the eye is low, fast motions tend to blur and the details are lost. If those motions are accompanied with a sound that somehow "follows" the outlines of the movement, the trajectory becomes much clearer. This phenomenon is used often in animation films, the typical example being a descending whistle sound accompanying an object's fall.[(1)] The function of these sounds is to clarify the spatial trajectory of the visual objects.

Weight

When seeking to map LMA's *Weight* effort factor to sound, it is important to ask whether there is any sonic equivalent; are there strong and light sounds? Since *Weight* can be defined as defined by the perceived amount of energy carried by a moving object, we could use the same energy model to talk about a sound's weight. It might be tempting to simply equate *Weight* with amplitude; however, a timbral definition might be more suitable. In this case, a *strong* sound would have a broad frequency spectrum, while sounds with a thinner harmonic content could be deemed *light*. For many cases, this link is satisfactory; a low piano note sounds stronger or heavier than a high flute sound. Some sounds are more problematic, though. A breath-like sound may have a very broadly distributed spectrum and still be perceived as *light*. To solve this problem, I propose the use of three criteria to assess the perceived *Weight* of a sound object. Firstly, as noted above, sounds with broader spectra tend to be perceived as *stronger*. This factor is more suited to compare two similar sounds. Secondly, sounds occupying a higher segment of the sound spectrum are generally perceived as being *lighter*. This is somewhat related to the first criterion, as higher-pitched sounds tend to have fewer partials, but has also a psychological component. An upward musical progression is usually associated by a certain light, buoyant quality. Thirdly, sounds with a coarser *grain*

(1) Chion, Michel *L'audio-vision* (pp. 102-104)

are often perceived as being stronger. *Grain* is one of the terms put forward by Pierre Schaeffer to qualify a sound's morphology.[1] He defines *grain* as the amount of small irregularities that affects the overall perception of a sound's texture. A pure sine tone has no such irregularities and thus has no *grain*. Introduce some amount of vibrato or tremolo and the *grain* factor increases. Sounds like that of a rasp that show a great amount of small, irregular variations in their perceived timbre are said to be very *grainy*. Introducing the *grain* criterion allows us to sort the problem of the breath sound, that may have a broad spectrum but is perceived as relatively light because of the regularity, i.e. low *grain*, of its texture.

The simple *one-to-one*, *empathic*, approach to *Weight* matching is obviously to equate changes in a movement's *Weight* to equivalent changes in the controlled sound's qualities. This is somewhat problematic, however, as moving objects' *Weight* factors are both somewhat difficult to accurately assess numerically and often relatively constant.

That is not to say that *Weight* relationships are unimportant in audio-visual pairings. Sound objects can play an important *materialisation* role. That is to say that through proper matching, sound objects can be used to influence the way a certain visual motion is perceived. The same movement, paired with two different sounds, one *light* and the other *strong* will tend to be perceived as having the qualities of the sound object. This happens, of course, within certain reasonable bounds. No matter how heavy, or strong, a sound is used to accompany the sight of a helium balloon rising, it is unlikely to have any effect on the perceived *Weight* of the balloon. This type of pairing can be said to be an extreme example of a contradictory approach. *Materialisation* works best with sounds that have very easily identifiable *Weight* qualities paired with rather ambiguous movements, *Weight*-wise.

(1) Schaeffer, Pierre *Traité des objets musicaux* (p. 548)

Flow

As has been explained in a previous section, the *Flow* effort factor is difficult to gauge numerically. That is why *Flow* does not make a particularly good source of control data. On the other hand, *Flow* terminology is fairly useful when describing sound-movement pairing strategies. *Flow* describes the amount of control that is being exerted over the moving object. *Flow* can thus be used to describe the quality of the control exerted not on the moving object, but by the moving object on the sound object. In a different approach, *Flow* can also be suggested by qualities of the sound.

A completely *bound* audio-visual pairing would be one where a given parameter is completely dependent on the state of the visual object. This approach is the easiest and generally corresponds to one-to-one or divergent mappings. A totally *free* pairing is one where the moving object exerts no influence on the sound parameter. Since a completely *free* relationship can also been seen as a lack of relationship, it is more useful to talk about *free* and *bound* in relative terms. A system can be implemented so that visual information only has a limited influence on sonic parameters; this would correspond to a *free* approach. This use of the *Flow* terminology differs with *empathy* in that *empathy* is a global, macroscopic, characteristic that extends to all levels of interpretation, be they *semantic*, *causal* or *morphological*, while *Flow* is strictly *morphological* and concerns itself with micro-organization.

Moment-Based Analysis of Shape

So far, I have presented a basic conceptual framework for controlling sound through visual movement. In order to implement this approach, however, some knowledge of computer vision techniques is necessary. In this section I will introduce a simple method of shape analysis. Of the many analysis techniques and algorithms available, I have chose moment-based analysis for its simplicity and because the measurements that can be calculated with it often correlate to readily identifiable visual features. By computing various moment-based shape descriptors, it becomes easy to express numerically many of the Laban Movement Analysis concepts introduced in previous sections of this thesis.

Definition

In moment-based analysis, the target binary shape is treated as the Cartesian representation of a two-dimensional discreet random function. Shape analysis is thus approached as a statistical problem.

The moment M_{pq} of the function is defined as:

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

In this situation, each pixel is treated as a sample point. For the purpose of binary image analysis, $f(x,y)$ is equal to 1 if the pixel is part of the shape, and 0 if not. x and y are the coordinate values of the pixel. p and q are positive integers. The order of a moment is the sum of p and q. (1)

Calculating the moment of an image for p=0 and q=0, in other words, M_{00}, is equivalent to counting the number of pixels in the shape. This is referred to as *mass*, or sometimes more intuitively, *area*. This is the simplest shape descriptor.

The values of M_{10} and M_{01} correspond to the sum of all x and y values, respectively. By dividing these two numbers by the *mass*, we obtain the mean horizontal and vertical values. If we treat these as a coordinate, we have the *centroid* or *centre of mass* of our shape. Knowing the *centroid* is of course useful as it allows us to know where our shape is located. As such, it can be used as a crude method of motion tracking. However, its greatest use is in allowing us to adjust pixel coordinates for other moment calculations. Instead of measuring the position of a pixel relative to a fixed point, usually the top left corner of the image, we can use the centroid for origin. Doing so has the advantage of rendering our calculations invariant to translation. Moments calculated about the mean in this fashion are said to be centralized. Centralized moments are written using the notation: μ_{pq}. Furthermore, we can use the mass to apply a normalizing function, making moments also invariant to scale; the same shape will yield the same values regardless of its position, or how big it is. Central normalized moments use the notation η_{pq}.

For the purpose of simple shape analysis and recognition, it is only necessary to calculate second and third order moments. That is to say that the sum of p and q will be equal to 2 or 3. I.e.

$$\eta_{20}, \eta_{02}, \eta_{11}, \eta_{21}, \eta_{12}, \eta_{30}, \eta_{03}$$

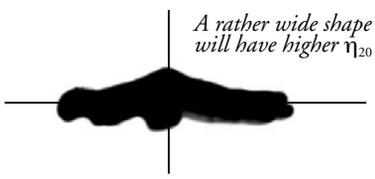
Moments as quantitative representations of qualitative features

Like the mass and centroid, second and

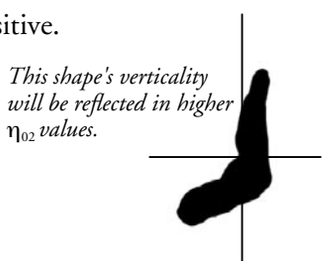
(1) Seul, Michael *Practical Algorithms for Image Analysis: Description, Examples, and Code* (p. 147)

third order moments can be used to derive useful information about the shape being analysed. Although moments are often used to calculate higher-level shape descriptors, they can be used "as-is" as indicators of certain visual features.

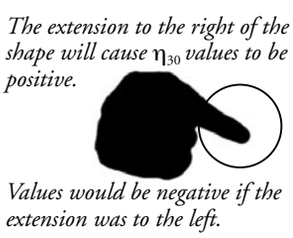
η_{20}, for instance, is a measure of horizontal distribution. Higher values correlate to “wide” shapes. Note that since this can be thought of as weight distribution, η_{20} is not necessarily related to the width of the shape. An example of this situation would be the case of the width being affected by a single pixel far from the centre.

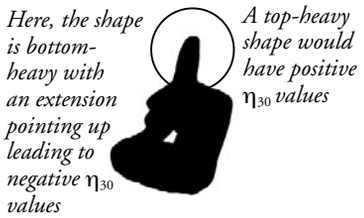


η_{02} corresponds to vertical distribution. Higher values mean that the shape is “tall”. η_{20} and η_{02} are negatively correlated, that is, when one increases, the other decreases. Owing to the power of two in the calculation, both values are always positive.

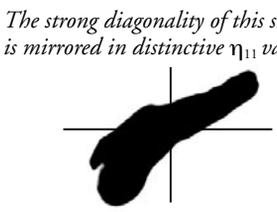


η_{30} and η_{03} are measures of asymmetry. Values close to 0 indicate high symmetry along the x-axis in the case of η_{30} and the y-axis for η_{03}. If a shape is bulkier to the right and rather outstretched to the left, η_{30} is going to be negative.





η_{11}, η_{21}, and η_{12}, are measures of skewness. η_{11}, specifically, is referred to as the correlation of the image. In practical terms this means that a shape that shows a distinct leftward lean will have a negative value, whereas a rightward lean will yield positive results. A shape with little or no skewness, like a square, will have η_{11} values close to 0.



Hu moment invariants

Central normalized moments may be invariant to scale and translation but they are very sensitive to rotation. In many cases, this is not necessarily a problem. In shape recognition, for instance, inverted shapes often carry different semantic values and the analysis method should reflect those differences. In some situations, however, this sensitivity can cause problems. Imagine, for instance, a setup where the camera is placed above the subject. It would be useful to make measurements that are invariant to scale, position *and* rotation.

Fortunately, such shape descriptors exist. From the list of seven second and third order moments, Ming Kei Hu derived a set of seven equations for computing descriptors that are invariant to rotation.[1] Those so-called *Hu moment invariants* are extremely useful for shape recognition but usually do not correlate

(1) Hu, Ming Kei *Visual Pattern Recognition by Moment Invariants*

to obvious qualitative features of the shape as moments do.

Other shape descriptors

Moments and Hu invariants can be used to compute other shape descriptors. These measures have the advantage of being often invariant to rotation and easily correlate with obvious shape characteristics.

Perimeter

The perimeter, in the context of a digital image, is the number of *edge pixels*. An edge pixel is defined as a pixel that is a) part of the shape and b) has at least one out of eight neighbours that is not part of the shape. The perimeter is obviously not invariant to scale but can be used to compute another descriptor: circularity.

Circularity

The definition of circularity is as follows:(1)

$$C = \frac{P^2}{4\pi M_{00}}$$

That is, the square of the perimeter divided by 4 times π times the *mass*. Circularity should equal 1 for a circle and tend towards zero for more complex or elongated shapes. In practice, however, owing to aliasing of digital images circles often do not neatly equal 1. Circularity is a good tool to discriminate between "tight" and "open" shapes, like, for instance, differentiating a clenched fist from an outstretched palm.

Elongation

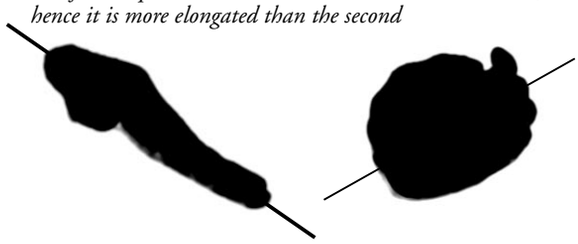
Elongation is a measure of how concentrated the shape pixels are along the main axis. Its formula is:(2)

$$E = \frac{\eta_{20}+\eta_{02}+\sqrt{(\eta_{20}-\eta_{02})^2+4\eta_{11}^2}}{\eta_{20}+\eta_{02}+\sqrt{(\eta_{20}-\eta_{02})^2+4\eta_{11}^2}}$$

A straight line has an elongation equal to 1, whereas a circle, the least elongated shape will yield a value close to 0. Elongation might appear to mirror circularity, however, since it does not rely on the perimeter it should be thought of as a measurement of distribution relative to the main axis. Elongation, as defined by this formula, is also referred to as *eccentricity*.

An alternate formula for the elongation measurement is actually the first of the Hu moment invariants: $\eta_{20} + \eta_{02}$. This simple measurement, unlike the previous formula, also factors in shape length and yields larger and larger values for shapes that are thin and long.

The first shape is more concentrated around its main axis, hence it is more elongated than the second



Orientation

The formula for computing the orientation of a shape is:(3)

$$\theta = 0.5 tan^{-1} \frac{2\eta_{11}}{\eta_{20}-\eta_{02}}$$

This is the angle of the main axis going through the centre of mass. There is some ambiguity for certain values, for instance 180° rotations but these can be adjusted for

(1) Parker, J.R. *Algorithms for Image Processing and Computer Vision* (p.292)
(2) Marshall, David *Visual Systems*
(3) Idem

using conditional terms. The measurement of orientation is, of course, much more accurate and meaningful for shapes that exhibit high elongation.

Shape transformations

Once various moment-based shape descriptors have been calculated, it is easy to use those measurements to assess LMA *Shape* parameters. (All references to moments from here on assume central normalized moments.)

Growing and *shrinking* motions will correspond to a change in *mass*, but all other parameters should remain fairly constant as the outline of the shape remains more or less the same.

Opening and *closing* changes can be assessed with the *circularity* measurement. As *circularity* values approach 1, the shape is *closing* - becoming more circle-like. Motion towards 0 will indicate *opening* or gradually more intricate outlines.

The *directionality* of shape changes will be reflected by differences in *orientation*. Furthermore, *asymmetry* measurements (η_{30} and η_{30}) can provide more sensitive information about the direction in which a shape might be deforming.

Effort factors

Space

Moment-based analysis can also be used as an easy method of performing trajectory analysis. Computing a numerical estimate for Laban's effort factor of *Space* is fairly straightforward. The position of the object is assessed using its centroids. This technique has the merit of being extremely cheap computation-wise and being very robust. In cases where the scene being analysed is too complex more elaborate algorithms may be used, such as Jean-Yves Bouguet's pyramidal implementation of the Lucas-Kanade optical flow estimation algorithm.[(1)] However, centroids correlate nicely to the perceived centre of the moving object and are ultimately the most suitable technique to use.

Trajectories are computed using a simple queue method. The positions of the last *n* centroids are drawn onto an *N* by *M* trajectory matrix *T*, where *N* and *M* are the dimensions of the image. Since it is safe to assume that the spatial transition between one centroid point and the following is a smooth one – i.e. points do not "jump" from one location to the next, linear interpolation is used to approximate the actual contour of the trajectory from the discreet points gathered. The number *n* specifies the size of the time window used. Care must be taken so that it is not too short, or too long. Typically, windows should last only a few seconds.

By calculating moments for the trajectory "shape," useful information can be readily obtained. The *directness* or *indirectness* of Laban's Space factor can be estimated by measuring the elongation of the trajectory. For this use, I favour the second, simpler equation ($\eta_{20} + \eta_{02}$). In this case, larger values will correspond to more direct motion, and lower numbers indicate indirect movement.

Weight

The *Weight* factor is more difficult to compute as it often depends on physical cues that are too subtle to be picked up by a computer vision-based system. For many non-human objects *Weight* is also often constant and does not provide useful control data.

A tentative measurement of *Weight* can nevertheless be ventured as a function of mass and trajectory curvature. Zhao proposes a complex neural network-based technique for

(1) Bouguet, Jean-Yves *Pyramidal Implementation of the Lucas Kanade Feature Tracker*

the estimation of the Laban effort factors.[1] I propose the following simple terms for the evaluation of *Weight*: a) The trajectory curvature of strong motions is less steep than that of light ones and b) light movements have a tendency to go against the main motion vector of the image. In practical terms, this second term translates as motions having a trajectory orientation that differs from those of other trajectories.

Time

The *sustained* and *sudden* qualities of effort are relatively easy to gather from the object tracking data. The distance from one sample to another is first calculated. Note that this need not be between two consecutive frames but at any time interval deemed suitable. The standard deviation of these distances is computed and is used to set a threshold value above which a movement is deemed to be sudden and below which it is sustained.

Flow

Flow is perhaps the most difficult parameter to express and estimate numerically. It might be construed as being related to velocity, however, this fails to account for contextual information that is important for the perception of *Flow*.

Shape recognition

Shape recognition is the act of matching a candidate object to a model previously recorded. Its use is obvious when it comes to using shape changes to trigger semantic relationships. There are a great many techniques available for this purpose here are two techniques that rely on moments and moment invariants. I developed the first one for a performance piece, while the second is a widely used technique.

Simple thresholding method

A data set of "positive images" is used to train the algorithm. "Positive images" are those of shapes that would qualify for positive identification of the target shape. Either moments or moment invariants are calculated for these shapes, depending on whether orientation is important. Those values are averaged and the standard deviation of each of the seven parameters (seven second and third-order moments or seven Hu invariants) is also computed. The same parameters are then calculated for the candidate shapes and their values compared to the model mean. A conditional statement of the type: if more than n moments are within a * standard deviation, then identification is positive, else it is negative. Here, n and a are used to adjust the tolerance of the system.

Care must be taken to select a good training set. Images that are too similar to each other will yield small standard deviation values and thus set the gates too narrow. On the other hand, a training set that exhibits too much variation will often lead to many false positives.

Mahalanobis metric

The Mahalanobis metric is a measure of the distance of a data vector to a given set. It is defined as follows:[2]

$$r^2 = (x - \bar{x})' C_x^{-1} (x - \bar{x})$$

Where r is the Mahalanobis metric, x is the candidate vector and C_x is the covariance matrix of the training data. The covariance matrix holds each terms covariance, that is, their tendency to vary together. By accounting not only for each moment's or invariant's tendency to vary independently but also relative to each other, the Mahalanobis metric often yields better results than the simple thresholding technique described above,

(1) Zhao, Liwei *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters* (p. 93)
(2) Duda, Richard O. *Pattern Classification (2nd Edition)* (pp. 35-36)

especially when using Hu invariants. The distance returned can be thresholded to provide Boolean identification. As with the other method, the results are highly dependant on the quality of the training set.

Computer Vision Tools for Artists

Even though consumer-level computers have only since recently been capable of real-time image analysis, the field of computer and machine vision is a relatively old one with much fundamental research being carried out in the 1960s and 70s. Perhaps owing to the stringent timing expectations demanded by the field, there have been few advanced image analysis systems available to artists. Here, I will present some of the more important ones currently available. The ideas put forward in this thesis are rather difficult to implement using most of those tools, however, which is why I developed the cv.jit package, which will be describe in greater detail at the end of this section.

BigEye

BigEye is a software system developed at STEIM[1] and which runs on Macintosh PowerPC computers. Up to sixteen different objects can be identified and tracked at the same time. Parameters used for object identifications are colour, size and intensity. The image frame is divided in a number of user-defined regions. These regions react to objects entering, exiting and moving within them. This information is sent out in the form of MIDI messages. BigEye also provides a scripting environment for carrying out tasks like conditional branching.

VNS

The Very Nervous System or VNS was created by artist David Rokeby[2] for an installation by the same name. In its first form, it consisted of both hardware and software components, to accommodate the slower computers of the time. The newest version, SoftVNS, is software-only and comes in the form of Max externals. The basic operation of the VNS system is frame subtraction. The pixel values from one image are subtracted from those in the previous and these differences are interpreted as movement. Like BigEye, the image frame can be divided in a number of zones and movement in each of these zones can be used to trigger various events. Although the earliest version, which dates back to 1982, was rather primitive, the newer SoftVNS 2 includes a number of more advanced algorithms, such as human head detection. The basic concept of VNS is still, however movement analysis: where and how fast is the image changing.

There are also two other systems that run within Max and are similar to BigEye and SoftVNS:

Cyclops[3] is a single Max object that divides video input in a rectangular grid of adjustable resolution. Each of the grid's cells are analysed in regard to motion and colour.

Êyês[4] is a more advanced suite of image analysis operators. It includes a number of filters and image segmentation tools and also supports zones in the manner of SoftVNS.

There is also the very flexible Jitter[5] suite for image and matrix operations that supports a number of analysis tasks such as histograms, FFT and simplistic colour tracking.

EyesWeb

EyesWeb is a full stand-alone software environment for the Windows environment developed at the *Laboratorio di Informatica Musicale* at the University of Genova. It is based in large part on OpenCV, an open-source

(1) STEIM *BigEye* http://www.steim.org/steim/bigeye.html
(2) Rokeby, David *Very Nervous System* http://homepage.mac.com/davidrokeby/vns.html
(3) Signer, David *Cyclops* http://www.cycling74.com/products/cyclops.html
(4) *Êyês* http://www.squishedeyeball.com
(5) *Jitter* http://www.cycling74.com/products/jitter.html

library of computer vision functions optimized for Intel processors, and provides a Max-like patcher approach to building programs.[1]

The primary objective of EyesWeb is to provide software tools for the real-time analysis of human motion. As such, it provides not only support for common computer vision tasks such as filtering, image segmentation and morphological transformations but also a number of human body-specific analysis modules.

EyeWeb's creators[2] also seek to provide a way to perform affective analysis of human gestures. They do this through the use of KANSEI information processing, which can evaluate non-logical data like personal tastes, and reference to dance and choreography-related theories of Laban and Eshkol and Wachman. In this regard, the goals of EyesWeb are similar to those set out in this essay. However, while focusing closely on human gesture allows them a greater level of precision and a higher level of analysis in their task, it shuts out some of the more general applications that are one of the great promises of computer vision.

cv.jit

It is this desire for a system that would provide: a) support for standard computer vision algorithms b) general shape analysis capabilities and c) good integration with sound synthesis software – in this case MaxMSP that led to my development of the *cv.jit* Jitter externals.[3] Although most of the capabilities of the cv.jit objects are mirrored in the EyesWeb software, working within the Max framework means both complete platform compatibility under Macintosh OS9 and OSX as well as WindowsXP. Furthermore, Max provides a much greater number of data processing tools and a generally more solid patcher environment than EyesWeb.

cv.jit objects are categorized into five categories: statistics, morphology, image segmentation, shape analysis and motion analysis.

Statistics

The statistics objects perform various tasks that are not specific to image analysis although they are often indispensable for these tasks. They include object for the calculation of the arithmetic mean of a sequence of matrices, as well as their variance and standard deviation. These measurements often prove quite valuable for background subtraction operations and the like. There are also objects that assist pattern recognition tasks by computing covariance matrices and Mahalanobis metrics.

Morphology

Morphological operators are typically used to prepare an image prior to analysis. The *erode* operation consists in assigning each pixel the minimum value of its neighbours. For binary images this can be restated as giving a pixel a value of 1 only if all of its neighbours also equal 1. It is so called because it makes image components smaller, as though eroded. The opposite of the erode operation is *dilate*. A *dilate* operation will assign a pixel the maximum value of its neighbours, or, for binary images assign a pixel a value of 1 if any of its neighbours equals 1, regardless of its original value. The tasks can be combined to perform *opening* (*erode* followed by *dilate*) and *closing* (*dilate* followed by *erode*).

Image segmentation

Image segmentation tasks currently supported are *flood fill* and *labelling*. *Flood filling* consists in isolating a single connected component. *Labelling* algorithms assign each individual connected component a unique

(1) *EyesWeb* http://www.eyesweb.org/
(2) Camurri, Antonio et al. *EyesWeb - Towards Gesture and Affect Recognition in Dance/Music Interactive Systems*
(3) Pelletier, Jean-Marc *cv.jit* http://www.iamas.ac.jp/~jovan02/cv/

value. The **cv.jit.label** object that performs this tasks also supports the removal of components smaller than a given threshold size and the automatic *labelling* of components with that component's *mass* rather than a sequential index.

Shape analysis

The shape analysis modules provide support for moment-based shape description outlined above. They include tools for computing *mass*, *centroids*, moments and moment invariants as well as secondary shape descriptors, such as *circularity* and *orientation*.

Motion analysis

Motion analysis object feature higher-level algorithms for motion tracking and analysis than available through the use of frame subtraction techniques or combination of image segmentation and centroids calculation. Two optical flow estimation objects are provided, one using the Lucas-Kanade[(1)] method and another the Horn-Schunck[(2)] method of estimation. (Optical flow measures the horizontal and vertical displacement of every pixel in an image. Although the methods used are too inaccurate for proper motion tracking they are invaluable when it comes to estimating the direction in which various objects in a scene move.) There are also a few objects to assist motion tracking, centred on an implementation of the pyramidal Lucas-Kanade algorithm. This allows for the accurate and robust tracking of up the 255 separate points in real-time.

(1) Lucas, B.D. and Kanade, T. *An Iterative Image Registration Technique with an Application to Stereo Vision*
(2) Horn, B.K. and Schunck, T. *Determining Optical Flow*

Kemuri-mai

Introduction

Kemuri-mai is a musical piece that aims to straddle the line between live performance and sound installation. A computer vision system is used to capture the various shape changes of the column of smoke rising from a stick of incense. These changes are used to control various sound parameters, thus resulting in a performer-less performance. Since the burning of incense takes place in a relatively narrow time frame – with a proper beginning, middle and end – the piece has a relationship with time that is more musical than it is related to the visual arts. This is a relatively site-specific work, as the sounds used are chosen partly in function of the space it is to be performed in. It is also an immersive experience. The incense, apart from providing subtle yet somehow stunning and ghostly visuals also infuses the room with its fragrance. *Kemuri-mai* also offers an example of how the concepts expounded in this thesis have been put in application. After briefly describing the technical aspects of the piece, I will address how audio-visual pairing strategies and computer vision techniques have been put to use.

Material set-up

The incense

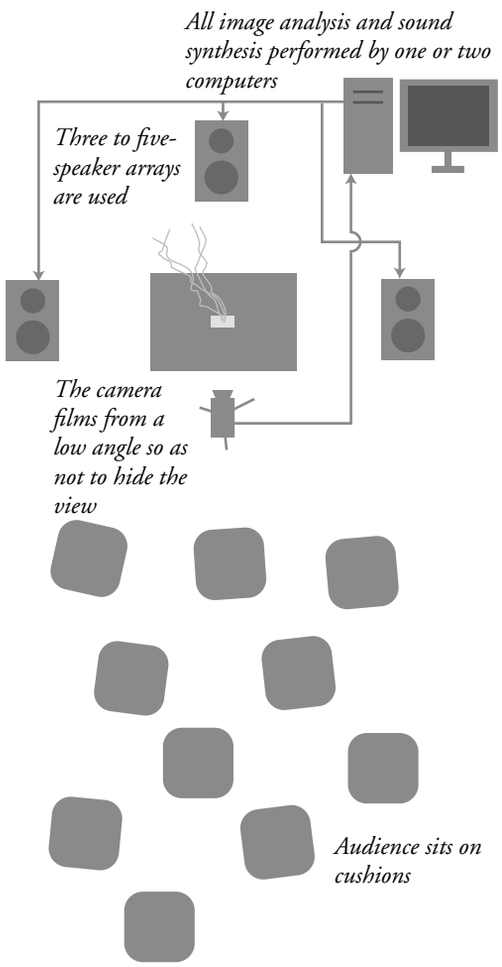
The choice of incense is important as not only various qualities and materials will provide different scents, from cheap soap-like odours to lush and complex spice fragrances, but will also have an important impact on the length it will take for the stick to burn and to a lesser extent on the thickness of the smoke. The length of the performance is decided by adjusting the length of the stick.

The hardware

The two main issues that have to be dealt with in this project, relative to hardware are light intensity and latency. In order for the smoke to be clearly visible, the piece must be performed in a very dark place. This means that the camera used must be very sensitive to low levels of light; the smoke is never going to be very bright. USB cameras do not provide anywhere near the level of accuracy required at such low light levels and cannot be used. IEEE 1394/Firewire/i-Link cameras provide acceptable image quality but typically cannot be zoomed, which limits their placement. The best approach, both in terms of image quality and latency combines the analogue output from a standard video camera and a frame-grabber device. Still, latencies of less than 100 ms are very difficult to achieve and need to be taken in consideration when designing the soundtrack.

Lighting is best achieved with very narrowly focused incandescent light sources, to avoid flooding the entire environment. It is important to have a very high contrast between the smoke and the background. This is important not only for the computer vision system but also for the spectators. If the circumstances allow it, preparing a black or dark background is ideal. The best lighting results are achieved by aiming the light source *towards* the audience, as the smoke will reflect almost no light back at its source. This is of course problematic; if the light is not placed properly it risks blinding the spectators. It is also possible to place the incense on a sheet of glass and light it from below. This method typically yields lesser contrast but the smoke can be seen from any direction. If an incandescent light source is placed too

close to the incense stick, the heat from the light bulb will cause vortexes that will agitate the smoke. This may be desirable as in certain environments the smoke can be quite static. By proper positioning of light bulbs, some amount of control can be exerted on the general behaviour of the smoke.



Software

The *Kemuri-mai* software runs entirely in MaxMSP. It is composed of two main modules: the image analysis module and the music synthesis module. In order to achieve better performance, it is sometimes necessary to use two different computers for these tasks and transmit the analysis data to the synthesis computer via MIDI.

The image analysis program uses Jitter, augmented with the capabilities of the *cv.jit* externals. There is no need for advanced filtering or colour recognition techniques in this piece and the smoke data can be identified using a simple intensity threshold approach. Any pixel whose intensity is greater than a given value are marked positively and passed on to the moment analysis module. Second and third order moments are then used to compute a number of relevant shape descriptors, which are then sent to the sound synthesis module.

The structure of the sound synthesis module is a lot freer. *Kemuri-mai* is a piece in constant progress and the soundtrack is created anew for each venue. Sample-playback and granular synthesis method often play a large role although classical additive and subtractive synthesis techniques are also used.

Shape analysis parameters

The shape descriptors calculated are based on the Laban Movement Analysis approached outlined in this thesis. As far as effort factors are concerned I decided to focus on *Time* and *Space*, as they are the two most relevant. A number of shape factors are also calculated.

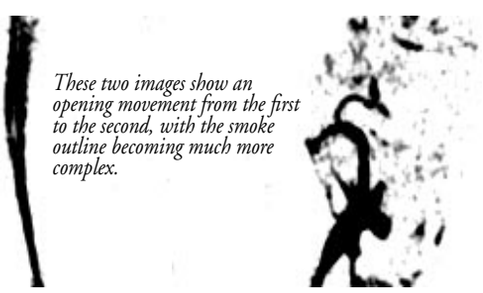
Shape Factors

The simplest shape descriptor used is that of *mass*. It is generally mapped to amplitude parameters so as to create a decrescendo effect as the smoke slowly vanishes in the last moments of the piece. Since the mass values change greatly depending on lighting conditions and threshold settings, I use a simple weighting of this and other parameters to ensure some amount of consistency.

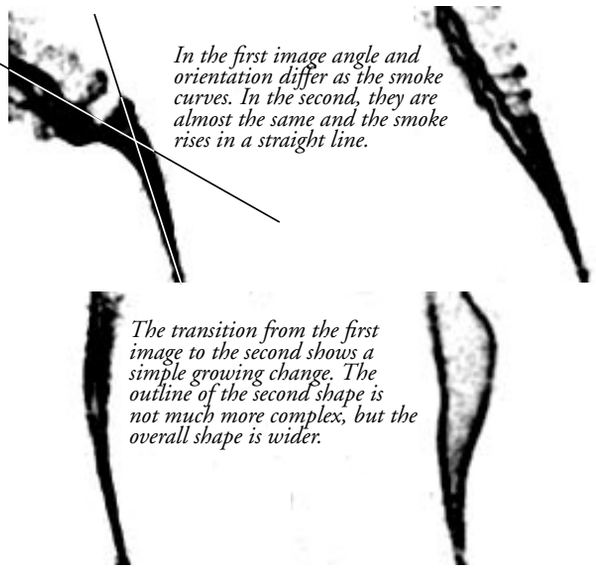
Two factors are used to measure movement. The first one, which is simply labelled *angle* is calculated by measuring the angle between the static base of the smoke column (a point I call *anchor*) and the *centroids*, which correspond roughly to the centre point. With this value we can tell by how much the column seems to be swinging and whether it leans to the

left or right. The second measurement is that of *orientation*, which is calculated using the formula given earlier. By comparing *angle* and *orientation*, we can have an idea of how much the column curves. If the smoke rises in a straight line, *angle* and *orientation* should be about the same, whereas the more curved it becomes, the more the two values will differ. If both values change together (i.e. are correlated), then the movement is qualified as *spokelike*, if they change relative to each other, the smoke is moving in an *arclike* fashion, to use LMA terminology.

I also calculate two *shape flow*-related descriptors. The smoke may sometimes rise in a thin line, whether curved or straight, or it can take a much more spread-out fan-like form. Movement from one state to another corresponds to *growing* and *shrinking* shape flows. Measuring this is easy with the second formula for *elongation* given previously. (The same formula as the first Hu invariant.) *Opening* and *closing* transformations, in this context correspond to changes in the shape's outline. A change from a state where the smoke rises in a compact column, be it wide or thin, to one that is curved and irregular is a form of *opening*, whereas the opposite is *closing*. To estimate this, I use *circularity* measurements as an *opening* transformation leads to a change in the mass/perimeter ratio.



These two images show an opening movement from the first to the second, with the smoke outline becoming much more complex.



In the first image angle and orientation differ as the smoke curves. In the second, they are almost the same and the smoke rises in a straight line.

The transition from the first image to the second shows a simple growing change. The outline of the second shape is not much more complex, but the overall shape is wider.

Effort factors

Using these shape descriptors, I also measure macro states that correspond to LMA's *Space* and *Time Effort factors*. If changes in *angle* are relatively slow and gradual I rank the motion as *direct*, spatially. Musically, this corresponds to a state I call *hei*, from the Japanese character for *peace, calm*. If, however, *angle* values change rapidly and irregularly, the movement is *indirect*, which is mapped to the musical theme of *ran*, or *disorder, war*. These two general states correspond to the most obvious general characteristics and most readily identifiable of the smoke patterns.

I also rank the various shape changes according to the *Time factor*. For each descriptor, I calculate the standard deviation. With this measurement, I set a certain threshold value for each parameter. If the change exceeds this value the movement is qualified as *sudden*, else it is deemed *sustained*. Again, these types of changes are very easy for the audience to discern. Sometimes, audience members might try to blow at the smoke to see what might happen. This is sure to result in a *sudden* parameter change and the music will change accordingly. However, if everyone starts blowing at the smoke continuously throughout the piece, the effect will be that the standard deviation value will inflate and the threshold for

a sudden change will be much higher, though the piece will be in a *ran* state.

It is highly possible that a particular performance of *Kemuri-mai* take place entirely in *hei* state with no sudden changes, just as it is possible that the piece never leave *ran* state. The system is extremely sensitive to small environmental changes that would normally be imperceptible. Sometimes, the turbulence caused by audience members breathing normally is enough to completely change the behaviour of the smoke compared to when the room is empty. Changes due to light bulb placement have already been mentioned. A light that is too close is sure to send the piece in perpetual *ran*.

Mapping approaches

As was already mentioned, Kemuri-mai is a piece in constant evolution, a perpetual *work-in-progress* of sorts. For this reason, the sound material changes from time to time. It is nevertheless possible to highlight some of the strategies used for the selection of sound material. It is important to note that the macroscopic musical structure is entirely dictated by the incense smoke's behaviour. Composition, in this context, consists in the choosing of sound material and the association of these sounds with various elements of the visual aspects of the piece. These choices are carried out based on criteria that fall in each of the three levels of interpretation highlighted in prior sections.

Causal pairing

Some level of causal linking is absolutely necessary for this piece to be successful. This is however difficult to achieve with the latencies of more than 100 ms I have to cope with. To solve this problem, I use one-to-one mapping of shape parameters only in *hei* mode. Furthermore, sound changes are made relatively smooth to mask the latency. This solution only works if the changes are relatively slow and gradual, which is why proper causal pairings are difficult to achieve in *ran* state. In the case of sudden parameters changes, sharp attacks may be used to highlight the drama. The delay between the shape change and the sound event may be acceptable and appear as a sort of echo. This depends greatly on the type of material used.

The semantic dimension

The *Kemuri-mai* soundtrack typically includes many recorded environmental sounds, in the tradition of *musique concrète*. These sounds are often left un-processed and recognisable which means they often carry a lot of semantic baggage. These sounds are carefully chosen for their relationship with both the incense and the place in which the piece is performed.

Incense as a cultural object has a fairly deep level of connotation, especially in *Kemuri-mai*'s country of creation, Japan. There is first of all an obvious religious subtext; even the fragrance alone is enough to evoke Buddhist temples. There is also a strong link between incense and death as it is used in funeral ceremonies and *kuyo*, memorial services. The incense's consumption also acts as a metaphor for ephemeral life. By changing the fragrance, the context in which incense is perceived can also be changed. For instance, it is easy to lend South Asian colours and a general exotic feel with certain smells. These various aspects are taken into consideration when choosing sounds in order to either bring out some elements of meaning, or to allow the incense to colour the interpretation of the sound material.

The performance environment is also very important. The likeliness of a certain sound being actually heard in the space, for instance, plays a big role in the selection. Various levels of meaning carried by the building, city, time of year, event context are all taken into

consideration.

Morphological mapping

Apart from the simple *Weight* to amplitude mapping previously alluded to, morphological pairings usually depend on the type of sound material used. Parameters like *angle* or *orientation* typically do not correlate very well with sonic parameters. In these cases more arbitrary pairings are chosen, such as *angle* to pitch. In the case of surround speaker systems, *angle* might be used to pan sounds around, though in practice, the *angle* is often too constant to yield interesting results. On the other hand, *opening/closing* and *growing/shrinking* transformations are useful to control *grain*-related parameters that are easy to change when using granular synthesis.

Sound material can also be used to *materialise* the smoke movement. *Light* sounds work well at emphasising this aspect of the smoke, whereas using *strong* sounds can sometimes create an interesting contrast.

Conclusion

Both the piece *Kemuri-mai* and the mapping strategies described in this thesis were developed from a desire to see more open-ended approaches to musical performance systems. By using computer vision, truly, anything visible can potentially be used to control musical or sound preformances. In front of the broad palette of possibilities, it is easy to feel lost, aesthetically and have recourse to simple and arbitrary strategies. It is important, however, to understand the various implications of the mapping decision one might take as a composer. While there exists some amount of research on what kind of useable information can be extracted from the environment, as well as inquiries in how musicians relate to performance systems,[1] I felt there was a certain lack in regard to how audio-visual pairs are perceived by th spectator. (This is of course excluding work done on motion picture sound.)[2] By combining elements of important theories of sound and movement, I have attempted to address the problem from this perspective.

Using computer vision in art (or any context) often presents important technical challenges. Small changes in environmental conditions like lighting can sometimes have disastrous effects on analysis programmes. However, the freedom of choice brought by, and the adaptability of vision-based systems makes up for these difficulties.

Future work will have to focus on sound and sight relationships in various settings to fully explore the potentials of the approach described herein. Some concepts might have to be adjusted, on the basis of practical experience. Easy to use software tools will have to be made available so as to make perception-based analysis of visual data available to more artists.

(1) Wanderley, M. et al. *Trends in Gestural Control of Music*
(2) Weis E. and Belton, J. *Film Sound*

References

Badler, N.I. *Temporal Scene Analysis: Conceptual Descriptions of Object Movement*. PhD thesis, Computer Science Department, University of Toronto, 1975

Bartenieff, I., Lewis, D. *Body Movement: Coping with the Environment.* Gordon and Breach Science Publishers, 1980

Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1983

Bouguet, J.-Y. *Pyramidal Implementation of the Lucas Kanade Feature Tracker.* from website of New York University, Media Research Laboratory: http://mrl.nyu.edu/~bregler/mocap2003/bouget00.pdf

Camurri, A., Poli, G.De, Leman, M., Volpe, G. *A Multi-layered Conceptual Framework for Expressive Gesture Applications*, Proc. Intl MOSART Workshop, Barcelona, Nov. 2001

Camurri, A., M. Ricchetti, R. Trocca et. al *EyesWeb - toward gesture and affect recognition in dance/music interactive systems.* Computer Music Journal, 24:1, pp. 57-69, Spring 2000

Chadabe, J. *The Limitations of Mapping as Structural Descriptive in Electronic Instruments.* Proceedings of the 2002 Conference on new Interfaces for Musical Expression, 2002
Siegel, W.m Jacobsen, J. *The Challenges of Interactive Dance: An Overview and Case Study.* Computer Music Journal, 22:4, pp.29-43, 1998

Chi, D., Costa, M., Zhao, L., and Badler, N. *The EMOTE model for effort and shape.* In Proceedings of SIGGRAPH 2000, pp. 173-182, 2000

Chion, M. *L'audio-vision 2e édition.* Nathan, 1990

Chion, M. *Guide des objets sonores: Pierre Schaeffer et la recherch musicale.* Institut National de l'Audio-Visuel & Buchet/Chastel, 1983

Chion, M. *Le son.* Nathan, 1998

Duda, R.O., Hart, P.E., Stork, D.G. *Pattern Classification 2nd Edition.* Wiley Interscience, 2000

Horn, B.K., Shunck, T. *Determining Optical Flow.* Artificial Intelligence 17:185--203, 1981

Hu, M.K. *Visual Pattern Recognition by Moment Invariants.* IRE Transactions on Information Theory, 8(2):179187, 1962

Laban, R. *The Mastery of Movement.* Plays, Inc., 1971

Laban, R., Lawrence, F.C. *Effort: Economy in Body Movement.* Plays, Inc., 1974

Lucas, B.D., Kanade, T. *An Iterative Image Registration Technique with an Application to Stereo Vision.* Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, pp. 674--679, 1981

Machover, T. *Hyperinstruments* from website of Massachusetts Institute of Technology, Media Laboratory: http://brainop.media.mit.edu/Archive/Hyperinstruments/index.html

Marshall, D. *Visual Systems*. from website of Cardiff University, Cardiff School of Computer Science: http://www.cs.cf.ac.uk/Dave/Vision_lecture/Vision_lecture_caller.html

Nyman, M. *Experimental Music: Cage and Beyond.* Cambridge University Press, 1999

Parker, J.R. *Algorithms for Image Processing and Computer Vision.* John Wiley & Sons, 1996

Peirce, C.S., (Hartshorne, C., Weiss, P. & Burks, A. Eds) *Collected Papers of Charles Sanders Peirce.* Thoemmes Pr., 1998

Rovan, J., Wanderley, M., Dubnov, S., and Depalle, P. *Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance.* KANSEI - The Technology of Emotion, AIMI International Workshop, Genova STUDY OF SPECTRO-TEMPORAL PARAMETERS IN MUSICAL PERFORMANCE FOR EXPRESSIVE INSTRUMENT

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

Saussure, F. *Cours de linguistique générale*. Librairie Payot, 1949

Schaeffer, P. *Traité des objets musicaux.* Éditions du Seuil, 1966

Truax, B. *Acoustic Communication 2nd Edition.* Ablex Publishing, 2001

Wanderley, M.M., Battier, M. (eds.) *Trends in Gestural Control of Music*. IRCAM - Centre Pompidou, 2000

Weis E., Belton, J. (eds.) *Film Sound*. Columbia University Press, 1985

Winkler, T. *Making Motion Musical: Gesture Mapping Strategies for Interactive Computer Music*. Proceedings of the 1995 International Computer Music Conference, 1995

Winkler, T. *Fusing Movement, Sound, and Video in Falling Up, and Interactive Dance/Theatre*

Production. Proceedings of the 2002 Conference on new Interfaces for Musical Expression, 2002

Winkler, T. *Composing Interactive Music: Techniques and Ideas Using Max.* MIT Press, 2001

Zhao, L. *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures*. PhD thesis, Computer and Information Science, Univ. of Pennsylvania, Philadelphia, PA, 2001.