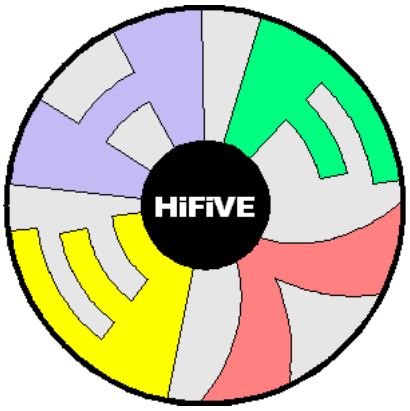

HiFiVE

An Audiotactile Vision-Substitution System

This poster describes "work-in-progress" on "HiFiVE" (Heard & Felt Visual Effects), an experimental vision-substitution system that uses verbally-orientated audiotactile methods to convey features of visual images to blind and deafblind people.

Introduction

There is often the need to convey general visual information to blind people. An existing approach is to use relief images e.g. tactile maps. While these are convenient for conveying unchanging two-dimensional images, the instantaneous production of vision substitution images is much more difficult to achieve. Devices can be devised that present other senses with information that includes aspects of sight, but other senses are not as powerful, or as able to comprehend such information.

The HiFiVE system aims to simulate the way that sighted people perceive visual features, rather than conveying raw optical measurements. The system highlights the features of visual images that are normally perceived categorically, and substitutes with coded sound effects (and tactile equivalents) : it simulates the instant recognition of properties and objects that occurs in visual perception, by using the near-instantaneous recognition of phoneme sounds that occurs in speech. The user can instantly understand the colours and objects that are present in an image by hearing coded phonetic sounds (and feeling the corresponding tactile/braille effects). The system also conveys shape and location, and may in future versions convey texture, distance and "change".

It is intended that the project will lead to a practical application. The HiFiVE system is designed to run on a personal computer. It uses standard equipment for sound output, and low-cost force-feedback devices for haptic input and output.

Coded phonetics

The HiFiVE system uses speech-like sounds, consisting of specific "coded phonetics" that can be rapidly interpreted in a categorical and linguistic way. These sounds convey the categorical properties of an image e.g. the colours, the distribution of those colours, recognised objects etc. People can easily recognise speech-like sounds and rapidly assign meaning to them. Most people are able to retain several such spoken words in their short-term memory, including "nonsense" words [3]. The effort needed to learn the coded phonetics is low.

Visual properties are presented to the user via groups of CV (Consonant-Vowel) syllables, each of which is assembled from a consonant and a vowel sound selected from a set of 16 consonant (C) and 16 vowel (V) sounds. The user can recognise the sounds instantaneously, in the same way as they can recognise language.

The approach is best illustrated by a simple example, shown in Figure 1. An image (A) is first reduced to 8 by 8 pixels (B). Then the pixels in each quarter of the image are set to one of two colours (C). Finally the image is presented via audio (D) and tactile (E) methods : for each quarter,

one CV syllable conveys the two colours, and two CV syllables convey the layout of those two colours, to the detail shown in the pixelated image (C).

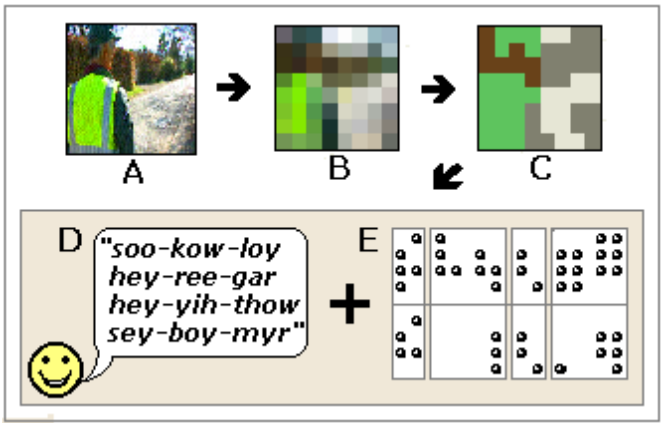

Figure 1. Diagram illustrating conversion of an image into coded phonetics and braille.

For the top left square of 4x4 pixels in the pixelated image (C), the CV syllable “soo” conveys the two colours, and the two CV syllables “kow-loy” present the layout of the two colours as 4x4 pixels. The whole image is conveyed by the four spoken words “soo-kow-loy, hey-ree-gar, hey-yih-thow, sey-boy-myr”, which completely describe the 8x8 pixels shown in (C). The user can control whether the system outputs audio and/or tactile effects to convey the colour combination and/or layout of colours. (Note that the conventions used in Figure 1 are likely to change before the system is completed.)

Other “Layouts” than the 8x8 format shown above can be used. For example, several “Layout” configurations based on allowing up to six 8-bit bytes of information being required to convey the configuration, i.e. requiring up to 6 braille cells or 6 CV syllables, are shown below. 6 braille cells can be read in one movement of a user's finger, and 6 CV syllables are straightforward for most people to remember short-term.

Figure 2 shows a 6-dot wide by 4-dot deep configuration - each of the three Layout braille cells/bytes has a corresponding braille cell/byte that specifies the 2-colour combinations to be used for it. The centre of the Layout is represented by a single unbroken cell.

Figure 3 shows a simple configuration - a single braille cell/byte specifies the 2-colour combinations to be used for it. The remaining cells can be used to present a 4x4, 6x4, 8x4 or 10x4 Layout by using 2, 3, 4 or 5 respectively of the remaining braille cells/bytes.

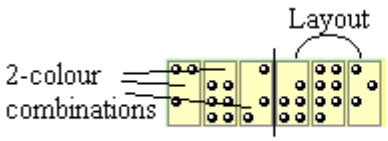

Figure 2

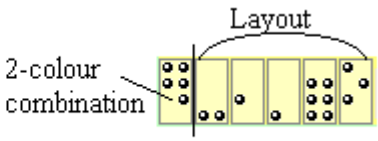

Figure 3

Many similar arrangements can be devised , e.g. a low-resolution Layout that has colours more precisely defined, 3-colour combination Layouts etc.

Tactile effects

The audio effects have tactile equivalents which can be presented using:- standard low-cost force-feedback devices to convey location and shape; and braille or other touch-based methods to convey the categorical properties.

As only 16 consonants and 16 vowel sounds are used, each CV syllable conveys one of 256 possible combinations (16 x 16). This corresponds to the number of different dot-patterns that can be presented on an 8-dot braille cell (refreshable 8-dot braille cells are available commercially [4]).

Figure 1 (E) shows one way in which the information given by the spoken sounds could also be conveyed via 12 braille cells.

A force-feedback joystick makes an effective pointing device with which to indicate areas of the image, as it can also be programmed to tend to position itself to one of a number of set positions, so that a "notchy" effect is felt as the joystick is moved, giving a tactile indication of location. A force-feedback joystick can also be moved by the system, pushing and pulling the user's hand and arm, both to convey shapes (by tracing them out), and to indicate a location within an image. Standard force feedback devices are currently used to present tactile effects.

Using both audio and tactile modalities allows the user to spread the load of information to suit their needs and abilities, and could be used by deafblind people. Having a degree of "redundancy" of information may result in less tiring usage [2].

Some of the other features of the HiFiVE system are described below:-

Audiotactile Tracers

The HiFiVE system can produce apparently-moving audio (and/or tactile) effects that trace out the shapes of features and identified objects within an image, by continuously changing the pitch and binaural positioning of the sounds (the "sound space" normally uses a high = high-pitch / low = low-pitch convention). These are known as "audiotactile shape tracers". In the tactile modality, tracer location and movement can be conveyed via a force-feedback device.

The audiotactile tracers can also move systematically through an area while outputting the properties of the corresponding parts of the image.

Moving effects are generally easier to mentally position than stationary ones.

Combotex

The fine detail of an area or entity can be conveyed via small, rapid fluctuations in the volume of the speech-sounds. These are referred to as "combotex" effects, as they combine the effects of small changes in brightness, colour, and distance, to give a single volume-conveyed "texture" effect. This simulates the effect found in vision whereby the overall properties of an area tend to be perceived categorically, and the minor variations in properties across it are perceived as a general texture. The user need not follow the precise detail conveyed by the combotex effects, but gets an impression of the general level of fine change occurring in an area.

Viewports

Sections of an image can be selected via a pointer, so that only those parts are conveyed, and at a higher resolution. These sections are known as "viewports", and the user can instruct a viewport to "zoom in" to higher levels of detail, as well as "zoom out" to convey a low-resolution representation of the whole image.

Viewports could be rectangular, hexagonal or "rounded" (circular or elliptical) and several viewports could be active at any moment. Viewports could be nested so that a "child" viewport moves within a "parent" viewport. One possible configuration would be to use nested viewports to simulate an eye's macula, fovea and/or areas of focal attention. (Only rectangular viewports are currently implemented.)

Audiotactile Entities

As well as conveying general visual features, the system attempts to simulate the way in which features and objects are perceived in vision. Conveying basic properties does not do much to

identify "entities", separate "figures" from the background, or assist with the other processes that occur naturally when people see things.

The simplest features are conveyed via shape-tracers, which can highlight identified shapes and features within a scene, emphasising the shape and layout of the feature or area.

Audiotactile Objects

“Audiotactile objects” are items in an image that have been identified to the extent that they can be presented as specific entities rather than being described in terms of their properties, shapes and features. They are signified by special CV syllables and braille patterns.

Initially audiotactile objects will mainly be used from within pre-processed images, but in the future automatic recognition of certain objects may be possible.

Image pre-processing

When completed, the system will be able to convey prepared programmes of material. Pre-processing allows a sighted designer to select features and areas of an image, and specify the most appropriate methods of conveying them.

Pre-processed information can be embedded in the image pixels using "steganography", so that the images can also be viewed by sighted people using standard equipment and software. Images and movie sequences prepared in this way can be transmitted through currently available media, e.g. via DVDs, the Internet or broadcasts, enabling pre-processed sequences to be embedded in otherwise standard video material.

(Audiotactile Entities, Objects and Image pre-processing have not yet been implemented.)

Activity-based processing

It is useful to be able to specify and store a set of filtering parameters / options that are to be used when users are engaged in a particular activity. The filters are performed in the following order :-

Basic filtering : This carries out the basic standard optical filtering of the image : cropping and stretching it; adjusting the brightness, contrast, hue and saturation; and any other standard optical filtering that is found to be useful for the activity in question.

It is useful to have a “de-masking” facility that removes the black masking that is often present in television broadcasts when a widescreen programme is presented on a 3:4 aspect ratio screen.

Area filtering : The Area filter uses standard optical processing techniques to isolate areas of common properties within the image.

The parameters for producing areas of common properties will vary according to activity.

Selection filtering : This decides which (if any) of the coloured areas isolated by the Area filter should be presented to the user. Its parameters include specifying the priority that should be given to each of several features, e.g. :-

- Colour shade
- Size (area and/or maximum & minimum dimensions).
- Closeness to an ideal shape.
- Closeness to an ideal orientation.
- Difference of colour shade from average shade (i.e. distinctiveness of colour).
- Separateness from edge of image.

  etc.

Each feature is assigned an importance, which can include “essential” i.e. areas failing to fall into a certain band of values for a feature can be excluded from consideration.

The system presents as many of the qualifying entities as it can in the time available for the current configuration, according to their score. Normally they will replace the standard Layouts in a panel in which they occur.

The system can be set to always analyse the whole image for entities, rather than the current viewport, and only present qualifying entities when they are found. This will generally be used for very precise/focussed activities e.g. “looking for a red ball”, and when the whole image is being inspected. If the required item is identified, then the system can move and resize the viewport to encompass it, and then present the viewport content in the normal manner.

The user will be able to rapidly change the Activity. For example, if they are watching a game of Snooker they could have Activities set up for seeking each of the ball colours, and switch between them when looking for different coloured balls.

Conveying vertices

Experiments show that emphasised vertices are particularly significant when conveying shapes via audiotactile methods : when people see items, vertices (such as the corners of a rectangle) produce a considerable effect in giving the impression of the shape, and it is important that this is reproduced in the audio and tactile modalities.

It is particularly important that vertices are highlighted when they are essential features of an entity but are not sharp angles, e.g. in an octagon.

HiFiVE momentarily stops the joystick movement to emphasise a vertex, and alters the sound volume.

For “live” images, the software that identifies the edge of objects could follow standard methods for identifying vertices, e.g. detecting when the direction of the perimeter changes at greater than a certain rate.

Colour shade allocation

HiFiVE makes use of the fact that when two shades are conveyed, they can be presented in either order; and the observation that most cultures tend towards classifying all colours into one of 11 colour categories, namely black, white, red, orange, yellow, green, blue, purple, brown, pink and grey [1]. These are known as “Basic Colour Categories”. The number of combinations of two of these 11 colours (presented in either order) is 55 (11 times 10 divided by 2) (e.g. “blue and pink”). If the shades “light grey” and “dark grey” are also included, but only included in combination with black, white or grey, then an additional 7 combinations are needed. Doing this allows “2-from-a-palette-of-5 levels” monochrome shade pairs to be conveyed. An additional code is assigned for “Polychrome” (i.e. no clear pair of colours), making a total of 63. This subset of the full 256 combinations of 16x16 CVs or 8-dot braille cells could be conveyed via 6-dot braille and allocated to the combinations that are easiest to recognise.

The remaining 192 codes can be assigned to extended shade ranges, which will include the 63 combinations already described i.e. the shade ranges can “share” the same single range of combinations. A convenient approach is to allocate 22 shades to the extended colour range; and 22 further CV combinations for presenting a single shade of the 22 colours (when a single shade predominates), making a total of 253.

The advantage with this approach is that several colour shade ranges can be used (e.g. “2-from-11-shades”, “2-from-22-shades”, “2-from-5-monochrome-shades”, single shades etc.), but the same continuous range of shades and combinations of shades is used. This means that the system can switch shade ranges without the user being particularly aware of it – e.g. from full colour to monochrome.

Specifying colour shade allocation

Different cultures may generally classify the colour shades into the “Basic Colour Categories” slightly differently, and the allocation of colours to shades is easily controllable by the user.

For particular tasks, users may wish to alter the extended colours, so that particular colours are more finely graduated.

Three-colour shade combinations
As well as the two-colour combinations and single-colour codes described above, the system could convey three-colour combinations via the 256 syllable or braille codes, as long as the colour shades are restricted to the 11 “Basic Colour Categories”. There are 165 3-colour combinations of 11 colours, 55 2-colour combinations, plus codes for the 11 single shades, totalling to 231 codes. The arrangement could include the additional shades “light grey” and “dark grey” for producing 5-level monochrome combinations.

Describing three-colour layouts is complex in both audio codes and braille, so it is better to confine any Layouts to conveying the arrangement of light and dark areas.

Speech coding

International codes - “KeliMopu” format
The coded phonetics “one of 16 consonants followed by one of 16 vowels” format described above have the advantage that they are language-independent. However they are based on the phonemes found in spoken English. The English language tends to have a higher number of basic phonemes when compared to other languages, which could make comprehension difficult for those for whom English is not their mother-tongue, especially as they will not be able to use the context to help with understanding, as would occur with normal speech.

As alternative format is to use 4 vowel sounds and 4 consonant sounds. For example vowels sounding close to the vowels found in the words “bed”, “bee”, “bog” and “boot” (when spoken with a Southern British Standard pronunciation) are found in most languages, and the consonants “K”, “L”, “M” and “P” are also found in most languages, allowing 256 double-syllable sounds to be produced, for example “keli” or “mopu”. A disadvantage of using this format is that the sounds take longer to speak, though the individual syllables can be presented slightly faster than the standard codes, as the vowel sounds and consonants are slightly shorter on average.

Simplified codes
For learners, the system could use shortened single-syllable versions of the English colour names, for names “Yell” for “yellow”, “Down” for “dark brown”, “Lown” for light brown etc. (Very short versions could also be used, e.g. “Yea”, “Dow” and “Low”.)

The disadvantages with this approach include that an additional set off sound samples is required; it is English-language-specific; it does not simplify the leaning of Layout codes; it takes longer to convey than fully-coded phonetics; and the users will probably want to learn the fully-coded versions anyway.

Force-feedback devices

Most force-feedback devices use Microsoft's “Direct-X” “Direct Input” protocol to interface with the computer. The devices found to be effective for use in connection with HiFiVE include Microsoft's “Sidewinder Force Feedback Pro” and “Sidewinder Force Feedback 2”; and Logitech's “Wingman Force Feedback Mouse”. The latter device has joystick emulation facilities, and is sufficiently powerful to pull the user's fingers to convey shapes, if it is loosely held.

Controlling force-feedback devices
The Wingman Force Feedback Mouse only has three control buttons, but a suitable communication protocol can be devised using multi-key sequences : up to 75 different signals can be given by pressing each button a maximum of one time but altering the order in which the three buttons are pressed down and released up i.e. so that the signal begins when one button is pressed, ends when all buttons are up, and each button can only be pressed once (some of the 75 combinations are quite complex and hard to use).

Two force-feedback joysticks could be used. Microsoft's. For example, the main joystick could be used primarily as the pointer, indicating the location and size of an entity (which may be very small); the other device, for example a force-feedback mouse (or second joystick) could be used to convey the shape of the entity, the tracer being expanded in size to better convey the detail of the shape.

When moving the viewport, as well as using the joystick's “hat”-switch to indicate the desired direction of movement, the system can be programmed to detect when the joystick has been pushed away from its current location by the user by over a certain distance. The system can then “snap” the viewport to the new location.

An alternative method of conveying categorical tactile effects
A low cost alternative method of conveying categorical information can be implemented by inducing coded pulses on a standard force-feedback device. Evenly-spaced pulses of two force levels are straightforward to interpret. These can be grouped into pairs of four pulses (separated by pauses), each group of four pulses allowing 16 possible combinations of strong and weak pulses, corresponding to a coded consonant (C) or vowel (V), with the two pairs (8 pulses) conveying the equivalent of a coded syllable or braille cell.

Pulse-based tactile coding is relatively slow compared to speech or braille, but may be useful for deafblind users who cannot read braille. It will not add to the hardware costs of a system that in any case uses a force-feedback device.

Summary

When fully implemented, the HiFiVE system will allow a continuum of visual features, from basic visual properties, to fully-identified objects, to be conveyed to blind and deafblind users. At the time of writing several of the features described above can be demonstrated, as well as “work-in-progress” on some of the others.

A final thought...

The system could be adapted to run in "reverse mode" : users could speak a version of the special coded sounds, and voice recognition software interpret the speech and frequency. Using "reverse mode", users could "paint" and modify an image using speech.

David Dewhurst
HiFiVE
www.HFVE.org