Converting SVG Images to Text and Speech

Vítor Carvalho[1] and Diamantino Freitas[2]

[1]University of Porto, Faculty of Engineering, Prof. Correia de Araújo Computer Centre,
Porto, Portugal
vitor@fe.up.pt

[2]University of Porto, Faculty of Engineering, Department of Electrical and Computer Engineering,
Porto, Portugal
dfreitas@fe.up.pt

ABSTRACT

There are contents with a strong visual component, e.g., in the engineering field, such as technical drawings, charts, diagrams, etc., virtually inaccessible for people with visual disabilities (blind, visually impaired, etc.). These contents are mostly vector based and can generally be found on the Web in various formats, but the recommended one is SVG.

The authors created an online application to convert SVG images (containing simple geometric figures of various sizes, filling colours and thickness and colour filling contours) in textual and spoken description by a speech synthesizer, client side based, without browser plugins. This application also allows the user to navigate with efficiency through the image description using four levels of detail and keyboard commands.

In this paper, the authors propose a novel method for image description based on the Gestalt theory, considering the cognitive load and, for the first time, providing users with visual impairments access to the full content of SVG images.

Application tests were carried out with 11 users (eight normally sighted, two blind and one amblyopic), comparing descriptions made by the application and by humans. The authors concluded that, for all users, there were improvements of 9%, using the application. Considering only the visually impaired, this figure rises to 18%.

1. INTRODUCTION

Visually impaired people represent 19% of world population. Of these, over 90% live in developing countries and over 18% have less than 50 years. Most of the visually impaired are not blind, suffering from various diseases such as glaucoma (which affects peripheral vision), age related macular degeneration (leading to loss of central vision), etc. [1].

Assuming that education is a value and a right for everyone and that all should have access with the best conditions [2], it is of interest to create an application that allows the translation of content with strong visual component (whose understanding is not directly transmitted through an alternative text), from the visual domain to other domains. One of these domains is hearing and, in this, one way of coding is by verbal description. Another possibility is to perform the conversion to Braille and use a standard reading ruler to read. The conversion to text precedes both conversions feeding oral description and representation in Braille.

In areas such as engineering, those contents may be technical drawings, graphs, charts and other complex documents, generally based on vector shapes produced by the respective editing programs. On the Web, the recommended vector format is SVG [3]. For this reason and at this stage, the authors carried out the development to this format.

Like everything that relates to usability and accessibility, this translation of the visual field to the auditory domain by means of oral description, benefits all users and not just those who have visual impairment. A relevant example is the ability to search images or image elements based on their respective content and characteristics, obtained from the text, which forms the basis for the oral description.

Thus, the authors set out creating an application that converts simple SVG vector images to textual description in natural language. The authors used open source and open access technologies to enhance its development. The application is Web based to reach more people and is independent of operating systems, browsers or specific plugins. It provides a synthesized voice to read the text description and, to control the cognitive load, it lets the user navigate through it, in an hierarchic way, using keyboard commands.

The authors present a review of the literature on this application in the next section followed, in section three, by the description of the proposed method. In section four, the preliminary assessment made on the application is presented and, in section five, the authors finalize with a conclusion.

2. LITERATURE SURVEY

The literature is rich in examples of raster image analysis but there is no analysis for SVG images in the desired manner, what means that there are algorithms to render SVG visually but not to do its automatic textual and natural spoken language description. The authors can advance some explanations for this:

- Raster images seem to be the preferred target of current investigation, perhaps because they are the

most abundant form on the Web, being sought after by many people;
- The SVG language is an XML dialect, thought to be sufficiently descriptive of its base visual content. However, the SVG description is not the convenient way to convey the description of the generated image. Therein lies the interest of this work.
- The authors also think that there is still insufficient awareness on providing accessible Web content.

This lack of related work led to the adaptation of some of the methods employed in raster images in the description of SVG images.

Ordonez et al. [4] developed and demonstrated automatic methods for image description using a large collection of captioned photos. They developed a technique that automatically collected one million images from Flickr, with filtered noise until the results of associated subtitles were visually relevant. This collection allowed dealing with the extremely difficult problem of generating relatively simple description using non-parametric methods and produced surprisingly effective results.

Recently, Yao et al. [5] presented in their article, image analysis for textual description (I2T), a structure that generates text descriptions of image and video content based on image study. The proposed I2T structure follows three steps:

1. They decompose the input images (or video frames) into their constituent visual patterns by an image analysis engine, in a similar spirit to analysing sentences in natural language.
2. Next, they convert the image analysis results into semantic representation as Web Ontology Language (OWL), which permits integration with general knowledge bases.
3. A text generation engine that converts the results of previous steps in readable text reports, semantically meaningful and subject to consultation.

The case studies demonstrate two automatic systems I2T: maritime and urban video surveillance system and an automatic system of real-time understanding of driving.

Castillo-Ortega et al. [6] present, in their article, a preliminary proposal to linguistic description of images. The base of approach is an hierarchical fuzzy image segmentation, a set of linguistic features describing each area and the diffuse spatial locations and relationships. The process is independent from the origin of these elements and provides a description with the characteristics of a synthesis, i.e., a brief and accurate description of the entire image. They disclose that this can provide a description of disjoint regions containing phrases appearing in different levels of detail.

Zhang et al. [7] holds that digital images are increasing worldwide. Thus, there is a growing interest in finding images in large collections or remote databases. In order to find an image, they must show or describe certain characteristics. The shape is an important visual feature of an image. Finding images using features related to the shape has attracted much attention. There are many techniques of representation and description in the literature. Their article classifies and reviews these important techniques. It also examines the implementation procedures for each technique, discussing its advantages and disadvantages, presents the results of some recent research and identifies promising techniques.

Although describing images is not the goal, the contribution given by Ferreira and Freitas [8] regarding the automatic reading of mathematical formulas in MathML contributed substantially in the preparation of this work. Both situations assume a common point, a document-based format of XML and the way the "navigation" occurs in a mathematical equation can find some parallelism in the textual description of an SVG image.

3. PROPOSED METHOD

Because automatic textual description of vector images seems to be little or not explored at all, the authors of this paper built an application with these objective.

3.1 Application Objectives

Since the beginning, the authors defined a set of goals that the application must fulfil:

- Convert SVG images to textual description in natural language;
- Use open source and freely available technologies to enhance its development;
- Use the Internet as its platform to reach the largest number of people;
- Be independent of operating systems, browsers or specific plugins;
- Provide a synthesized voice to read description, allowing navigation through it.

The application is able to describe the contents of SVG images with respect to their shape, colour and spatial location. It also describes the perceptual organization of the constituent elements of the image, according to the Gestalt theory, focusing on aspects such as the occurrence of symmetries, alignments and group formation.

3.1.1 Shape

The elemental form is a basic visual feature to describe the content of an image [7] and is an important property for the perceptual recognition of objects or elements and classification of images [9]. However, the representation of shape and its description is a difficult task in raster images [7]. Given the descriptive and parametric nature of SVG, the task is easier for the application.

Falomir et al. [10] present descriptions of some shape attributes, such as comparing the lengths used by the application.

Through the interpretation of the SVG code, the application can recognize a relevant set of geometric shapes in the image, namely:

- Rectangles (<rect>), also identifying if the rectangle is a square, using the method of Falomir et al. [10] for length evaluation;
- Circles (<circle>);
- Ellipses (<ellipse>);
- Polygons (<polygon>), also identifying if the polygons refer to triangles, lozenges or polygons of other type.

3.1.2 Spatial Location

Falomir et al. [10] proposes, following other models, a description of absolute and relative orientation of objects with the observer based on the division of space in eight directions surrounding each object: left, front-left, front, front-right, right, back-right, back, left-back. This work considered another approach. The authors divided the analysed image into nine equal parts (Figure 1): upper left, upper centre, upper right, centre left, centre, centre right, lower left, lower centre and lower right.

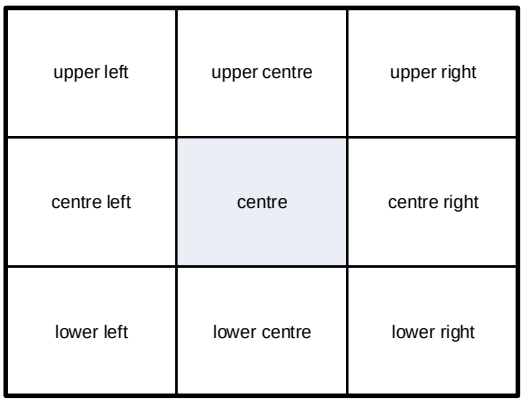

Figure 1. Division into Nine Equal Parts of the Analysed Image.

If the point belongs to the central part of the image, the authors perform a second and thinner correspondence (Figure 2): a little to the left, a little above, a little to the right and a little below the middle.

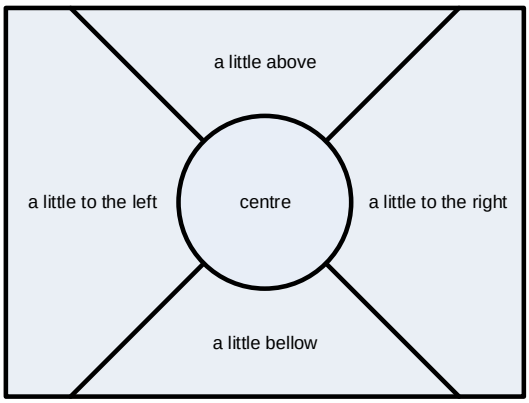

Figure 2. Central Part of the Analysed Image: Division into Five Areas.

3.1.3 Colour

The colour names are linguistic labels given by humans. They are used routinely and similarly with no effort to describe the world. Identifying the colour by its name is a method of communication that all people understand [11]. The fields of visual psychology, anthropology and linguistics studied colour naming.

The colour information is critical in applications such as art, fashion, product design, advertising, film production and printing [12]. Mojsilovic [11] seems to have a similar opinion when she says that although the naming of colours is one of the most common visual tasks, it did not receive significant attention by engineers.

Today, with the emergence of visual technologies, sophisticated interfaces with the user and man-machine interactions, the ability to appoint individual colours, pointing out objects of a certain colour and convey the impression of dithering, becomes an increasingly important task [11].

The advantages of naming colours have to do with image search, automatic image labelling, assisting individuals who are colour blind and human-computer linguistic interaction [13]. Everyone can benefit from automated methods to describe and recognize colour information [12].

Falomir et al. [10], in their qualitative description of colour, begin to translate it from RGB (Red, Green, and Blue) system to HSL (Hue, Saturation and Lightness), claiming to be most suitable for colour nomination by ranges of values. Guberman et al. [14] seem to have same opinion when they say that it is a more fruitful approach to colour from the human point of view, i.e., based on the luminosity, gamma and saturation instead of the amount of additive primary colours like red, green or blue.

Falomir et al. [10] present a model for the qualitative colour description which separates only Lightness dependent colours (black, dark grey, grey, light grey and white) of the remaining specified colours (red, yellow, green, turquoise, blue, purple and pink). The latter still have name variations adding “pale”, “light” and “dark” adjectives.

This paper made a few adjustments to the Falomir et al. proposed model [10], introducing the brown, orange and yellow colours. The authors translate variations of colours dependent from other components besides lightness by the following adjectives: “pastel”, “light”, “whitish” and “dark”.

3.1.4 Perceptual Organization

Wagemans et al. [15] advocate that grouping is the most associated with visual perceptual organization phenomenon. This effect results when some elements of the visual field appear to be more together. This also happens when the clues are weak and disparate. The human brain seems to have the ability to combine them synergistically in order to form strong evidence of grouping.

Some psychophysical and computational studies about groups, using carefully built stimuli [16], allowed

quantifying some grouping principles used in this work, based on the Gestalt theory:

- **Proximity –** grouping between two elements increases as these elements are closer to each other; conversely, the strength of proximity grouping exponentially decays with increasing distance between the elements [16].
- **Shape perception** – symmetry, parallelism [15].

With only two elements in the image, the application perceives horizontal and vertical relations of symmetry relative to the central vertical and central horizontal axis of the image. For this purpose, there is a function that receives the coordinates of the central point of each image element to determine if the x and y coordinates are almost the same.

If there are more than two elements in the image, the application performs a correlation analysis between the coordinates of the centre points looking for linearity by the linear correlation coefficient R (Equation 1) [17].

$$R = \frac{\sum x \times y}{\sqrt{(\sum x^2) \times (\sum y^2)}}; x = x_i - \bar{x}; y = y_i - \bar{y}; \quad (1)$$

If the coefficient R is greater than 0.9 (positive linear correlation) or less than -0.9 (negative linear correlation), the elements are considered to be aligned.

Another analysis of this level is the formation of groups.

For this, the authors developed an algorithm, based on a nearest neighbour analysis (Equation 2) [18].

$$R_n = \frac{\bar{D}(Obs)}{0{,}5 \times \sqrt{\frac{a}{n}}} \quad (2)$$

In this formula, the numerator is the average distance of the observed nearest neighbour and in the denominator, under the radicand, the study area above the total number of points. This formula gives a value between 0 and 2.15. It takes the value 0 if the points are close together, 1.0 if randomly arranged and 2.15 if arranged regularly. Although the authors consider incorporating the formula, as is, in a future version of the application, for now the decision was to implement an empirical formula to return groups of elements in the image (even if each group had few elements).

Thus, the proposed algorithm measures the distance of each element in the image to the other elements. If this distance is less than one quarter of the maximum diagonal image, it is considered that these elements form a group. The algorithm ensures that each pair of elements appears only once.

3.1.5 Description and Navigation Synthesis

The image description and navigation through the description suits the notions considered as the basis of the Gestalt theory, presented by Wagemans et al. [16].

- **Holism** – the perceptual experiences are intrinsically holistic and organized by rejecting atomism and associationism, as well as any summative approach. Whatever parts (properties, elements) are perceived holistically and not in a separate or independent manner. However, persons perceive shape and colour separately.
- **Emergency** – emergent properties and superiority of configuration. Emergent properties belong to the whole and not the individual parts (e.g., the density of a forest applies to the whole forest and not to a single tree).
- **Configuration superiority** – persons perceive the parts after the whole (using the same example, an observer perceives the existence of a forest before focusing on the trees that compose it).
- **Global precedence** – processing happens from the global structures to the analysis of local properties. The persons process first the overall properties of a visual object, followed by the analysis of the local properties.
- **Primacy of the whole** – the properties of the whole cannot derive from the properties of its constituents. They are born of inter-party relations: symmetry, regularity, closing, etc.

Thus, the authors picked four levels of detail, so that the description can achieve the accuracy required by the growing complexity and level of detail while providing precise and cognitively correct brief descriptions.

The first level focuses on the aspect of image, background colour, number of elements, title and the SVG base description.

The second level breaks down the elements regarding the form, indicating how many elements of each type exist. It also concerns some aspects of the Gestalt theory, regarding the picture as a whole, evaluating the existence of alignments, symmetries and group formation.

The third level is concerned with elements description, indicating its type, approximate dimensions in relation to the image size, approximate location relative to image, the name of the fill colour and, for the perimeter, the name of its colour and approximate thickness.

The fourth level is similar to the third but introduces a deeper technical description, in which all values are numeric, in order to disambiguate any approach or names given in the third level.

3.2 Used Technologies

The SVG interpreter is coded in PHP, using the SimpleXMLElement. From the parsing of the SVG file, the authors created methods for naming colour, location of elements in the image, size comparison and Gestalt analysis.

The phrase builder for the construction of the textual description takes all these methods and prepares parcels of text for each level of detail, element, shape and element location.

The interpretation of keyboard commands regarding navigation through the description is JavaScript, which passes to the voice synthesizer the blocks of text to read.

The speech synthesizer used in the tested version is a routine called meSpeak [19], client side type, working on JavaScript and JSON.

3.3 Limitations

Currently there are some limitations in this application.

- Unsupported SVG tags:
  - `<line>`;
  - `<polyline>`;
  - `<g>`;
  - `<text>`;
- Unsupported content:
  - CSS styling;
  - Animations;
  - Dynamic behaviours;

Because it is an application with little maturity, error handling is not yet implemented.

3.4 General Using Principle

The steps that the user must perform to convert an SVG image to textual description, receive and use the result is:

- Access the application site (currently in Portuguese), available on: http://paginas.fe.up.pt/~vitor/svg2desc/
- Submit the SVG file to convert to textual description in natural language;
- Get back the description in the form of synthesized speech and written text;
- Navigate through the description by keyboard commands.

3.5 Application Interfaces

3.5.1 Main Page

The initial interface (Figure 3) shows, after the main title, the instructions on how to interact with the application:

- Choose the desired SVG file;
- Select the most convenient options;
- Click on “Convert to SVG textual description” button.

Immediately after the instruction is a label with "SVG File" and ahead of this, a button that allows the localization of the desired SVG file (Figure 4).

Before the user clicks the button to convert the SVG, the application shows three options. The most important are:

- Display the text of the simple description (Figure 6) - in addition to speech synthesis, the application displays the text of levels 1, 2 and 3 (see in chapter 3.1.5 the meaning of these levels);
- Display the technical text of the description (Figure 7) - in addition to text-to-speech, the application displays the text of level 4 (see in chapter 3.1.5 the meaning of this level);

To start the process of converting SVG to textual description in natural language the user must click the "Convert to SVG Textual description" (Converter SVG para Descrição Textual) button.

Figure 3. Application Main Interface (in Portuguese).

Figure 4. SVG File Selection.

3.5.2 SVG Description

The SVG description interface (Figure 5) could not be simpler.

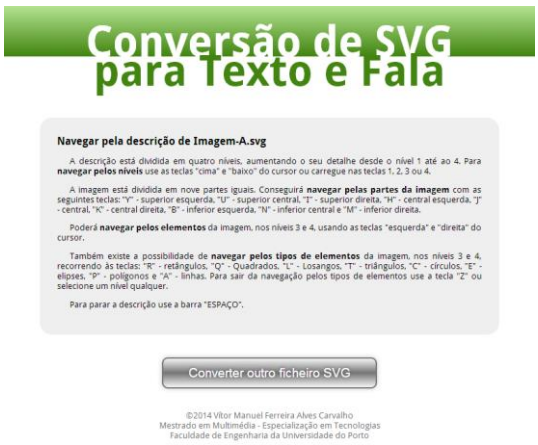

Figure 5. SVG Description Interface (in Portuguese).

Soon after the page title, instructions on how to navigate through the SVG description are included:

- The description divides into four levels, increasing its detail from level 1 to level 4. To **navigate through the levels**, the "up" and "down" cursor keys or numeric keys "1", "2", "3" and "4" must be used;
- The image divides into nine equal parts. To navigate through the parts of the image use the keys: "Y" - upper left, "U" - upper centre, "I" - upper right, "H" - central left, "J" - central, "K"- central right, "B" - lower left, "N"- lower centre and "M" - lower right.
- To **navigate through the image elements** in levels 3 and 4, use the "left" and "right" cursor keys.
- There is also the possibility to **navigate through element type**, at levels 3 and 4, using the keys: "R" - rectangles, "Q" - Squares, "L" - Lozenges, "T" - triangles, "C" - circles, "E" - ellipses, "P" - polygons and "A" - lines. To exit the navigation by type of element use the "Z" key or select any level.
- To stop the description use the "SPACE" bar.

The name of the loaded SVG file appears as a subtitle of this interface allowing the user to know in what context the description is.

Once the SVG is loaded and interpreted, the synthesized voice announces "Ready to make the description of the image." That is the cue for the user that the application is ready to receive keyboard commands.

Depending on the options that the user chooses the initial interface, the interface description of SVG can take several aspects.

The first aspect (Figure 6) results from selecting the option "Display the text of the simple description". This text appears just below the navigation instructions in a box with green background.

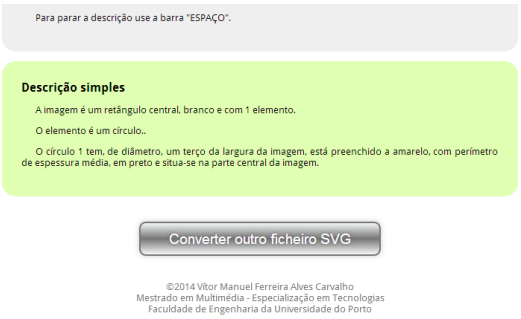

Figure 6. SVG Description Interface with Simple Description (in Portuguese).

The second aspect (Figure 7) results from selecting the option "Display the technical text of the description." This text appears just below the navigation instructions in a box with salmon background. If the user selects both options, the two boxes appear.

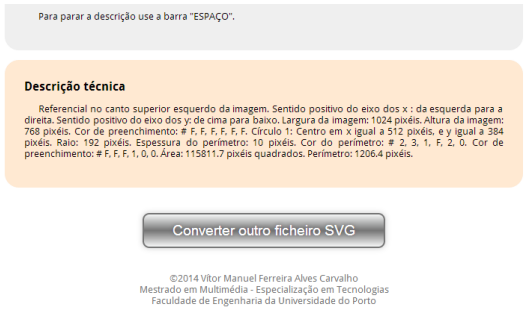

Figure 7. SVG Description Interface with Technical Description (in Portuguese).

4. EVALUATION

4.1 Test Images

To test the application the authors chose eight images (Figure 8) created with Adobe ® Illustrator ® CS6 software, thought to represent some of the problems of analysis and interpretation, including the Gestalt theory, spatial visualization, semantic and cognitive load.

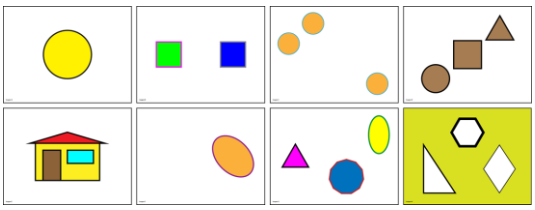

Figure 8. Test Images.

Each image has an associated set of attributes that users need to memorize:

- Image background colour;
- For each of its constituent elements:
  - Location;
  - Size;
  - Shape;
  - Fill colour;
  - Perimeter thickness;
  - Perimeter colour.

4.2 Example of a Description made by the application

The application describes the first image in the top left corner of Figure 8 as (translated from Portuguese):

Level 1:
The image is a horizontal rectangle, white and with one element.
Level 2:
The element is a circle.
Level 3:
The circle 1 has a diameter of one-third the length of the image, filled in yellow, with a medium thickness black perimeter and is located in the centre of the image.

4.3 Test Images Description by Expert Users

First, three expert users from University of Porto described the images (Table 1).

User	Degree	Area
UE1	PhD	Fine Arts / Web Design
UE2	PhD	Engineering / Computer Graphics
UE3	MSc	Engineering / Technical Drawings

Table 1. Expert Users for Image Descriptions

On average, each image took the experts four minutes to describe.

4.4 Test Subjects Without Visual Impairment

The authors picked up eight persons (Table 2), college graduates at University of Porto in areas where SVG images may be relevant.

User	Degree	Area
U1	MSc	Geography
U2	*Graduate*	Informatics
U3	*Graduate*	Mathematics
U4	*Graduate*	Physics / Informatics
U5	*Graduate*	Informatics
U6	*Graduate*	Computer Engineering
U7	*Graduate*	Informatics
U8	*Graduate*	Informatics

Table 2. Tested Users without Visual Impairment

In the first part of the test the participants were asked to read the description of four images on paper (made by one of the expert users), taking the desired time to understand the image depicted. As soon as they completed the reading of an image description, the description was removed and they were asked to draw on paper the image corresponding to the description using felt-tip pens of different colours (e.g. in Figure 9).

In the second part of the test, after a period of clarification on the operation and navigating through the description using the application, they were asked to hear the description of four images, taking the desired time to understand the image depicted. As soon as they completed the image description hearing, the application was withdrawn and they were asked to draw on paper the image corresponding to the description using felt-tip pens of various colours (e.g. in Figure 9).

The authors had the care to distribute randomly the expert users' descriptions, always having one male and one female as a reader.

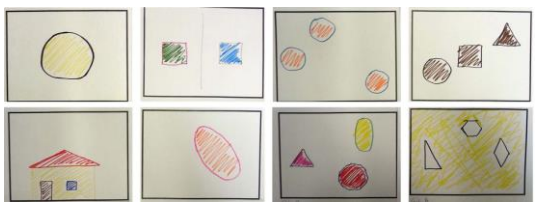

Figure 9. User U4 Drawings

The authors gave the images in random order for each pair of the previously referred users. Four image descriptions to read on paper and four to hear in the application. However, if the first element of the pair read the description of the image, the second element of the pair heard the description of the image in the application and so on. The user U6 heard all the images in the application in a random sequence.

4.5 Tested Users With Visual Impairment

The authors picked up three persons (Table 3), college graduates at University of Porto in areas where SVG images may be relevant.

User	Degree	Area	Impairment
U9	*Graduate*	Information Science	Blind
U10	MSc	Literature	Blind
U11	*Graduate*	History	Amblyopic

Table 3. Tested Users with Visual Impairment

In the first part of the test, after a period of clarification on the operation and navigation in the description using the application, users were asked to hear the description of four images, taking the needed time to understand the image depicted. Once they concluded hearing the description of the image, the application was withdrawn and they were asked to compose, using various geometric sponge figures, the image corresponding to the description (e.g. in Figure 10).

In the second part of the test the users were asked to read the expert users' description of four images (stored, each one of them, in a text file) using their usual screen reader and taking the desired time to understand the image depicted. As soon as they completed the reading of the image description, the computer was removed and they were asked to compose, using various geometric sponge figures, the image corresponding to the description (e.g. in Figure 10).

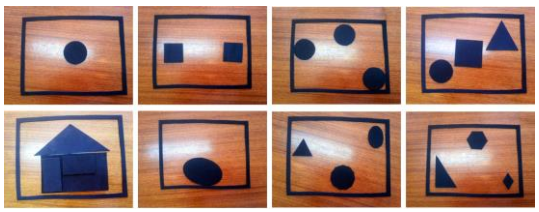

Figure 10. Example Compositions (User U9).

4.6 Score from assimilated descriptions

The authors did the drawings and compositions analysis not for its artistic quality, but rather for what users wanted to represent. The authors took notes during user testing and, jointly, an exempt person evaluated the drawings in a later stage of the data collection. The authors sought for the accordance of the aspects for each image in the qualitative characteristics described in paragraph 4.1.

For a correct aspect, one point was assigned. Being incorrect, zero points. At the end of each image and for each user, the authors found the average of all the attributes and multiplied by 100 to obtain a percentage termed "perfection".

4.7 Conclusion from the tests

The first conclusion to draw is that the description of images by expert users is a time consuming task, with an average of four minutes for each image.

The result of the expert users' descriptions was not ideal, because in this evaluation work more than half of the end users pointed out that the descriptions were not homogeneous and used ambiguous terms to describe the images. In fact, the average "perfection" (see paragraph 4.5) of all users who read the expert users' descriptions was 70%. The average "perfection" for users who heard the descriptions in the application was 79% (9% more than reading on paper). Considering only the visual impaired users, the average stands at 87% using the application and 70% using the screen reader (expert users' descriptions), a fall of 17%.

Taking into account all tested users, hearing a description of the image in the application is, on average, 57 seconds slower than the equivalent reading on paper or screen reader. Considering only the blind users, hearing the description of the image in the application is on average 21 seconds faster than hearing on their own screen reader.

On the other hand, taking into account all users, there was an improvement in "perfection" of 7% when they heard the description on the application rather than reading on paper or by a screen reader. Users who decreased their performance when using the application, worsened on average 1%. Users who improved their results when using the application, improved on average 15%.

Given the above, the authors consider that the application is an asset to describe SVG images, benefiting general users and even more the visually impaired.

5. CONCLUSIONS

The authors built an online application that, from an SVG image, automatically makes its textual description with some communication effectiveness.

It was possible to integrate text-to-speech considering pragmatic and improving its pronunciation of words.

Whenever possible, the application analyses the images in search for organizational structures of its elements, based on the Gestalt theory, considering symmetry, alignment and group formation.

The application was tested with a representative set of users and the authors found that the best results were achieved using the developed application. Taken all users, there were improvements of 9%. If the analysis focuses only on users with visual impairments, this figure rises to 18%.

A larger population will further test to build statistical proof.

In this work there is innovation regarding the transmission of information contained in SVG images for people with visual impairments by an auditory pathway. The authors addressed sensitive issues such as the implementation of descriptions regarding cognitive loads.

The application provides, for the first time, access to the full contents of SVG images to users with visual impairments using a new paradigm to navigate through the image description, based on level of detail, content location and shape type.

REFERENCES

[1] T. Pun, P. Roth, G. Bologna, K. Moustakas e D. Tzovaras, "Image and video processing for visually handicapped people," *Journal on Image and Video Processing*, vol. 2007, p. 12, 2007.

[2] SAEDUP, "Serviço de Apoio ao Estudante com Deficiência da UP" Universidade do Porto, June 4th, 2013. [Online]. Available: http://sdi.letras.up.pt/default.aspx?pg=saedup02.ascx&m=11. [Accessed on June 17th, 2013].

[3] W3C, "Scalable Vector Graphics (SVG) 1.1 (Second Edition)," W3C, August 16th 2011. [Online]. Available: http://www.w3.org/TR/2011/REC-SVG11-20110816/. [Accessed on February 5th, 2014].

[4] V. Ordonez, G. Kulkarni e T. L. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," in *Neural Information Processing Systems Foundation*, Granada, 2011.

[5] B. Z. Yao, X. Yang, L. Lin, M. W. Lee e S.-C. Zhu, "I2T: Image Parsing to Text Description," *Proceedings of the IEEE*, vol. 98, nº 8, pp. 1485-1508, 2010.

[6] R. Castillo-Ortega, J. Chamorro-Martínez e N. Marín, "Describing Images Via Linguistic Features and Hierarchical Segmentation," in *2010 IEEE International Conference on Fuzzy Systems (FUZZ)*, Barcelona, 2010.

[7] D. Zhang e G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, pp. 1-19, 2004.

[8] H. Ferreira e D. Freitas, "Leitura Automática de Fórmulas Matemáticas". Master of Science Dissertation on Informatics Engineering, Porto: Faculty of Engineering of the University of Porto, 2005.

[9] B. G. Prasad, K. K. Biswas e S. K. Gupta, "Region-based image retrieval using integrated color, shape, and location index," *Computer Vision and Image Understanding*, vol. 94, pp. 193-233, 2004.

[10] Z. Falomir, E. Jiménez-Ruiz, M. T. Escrig e L. Museros, "Describing Images Using Qualitative Models and Description Logics," *Spatial Cognition & Computation: An Interdisciplinary Journal*, vol. 11, nº 1, pp. 45-74, 2011.

[11] A. Mojsilovic, "A Computational Model for Color Naming and Describing Color Composition of

Images," *IEEE Transactions on Image Processing*, vol. 14, nº 5, pp. 690-699, 2005.

[12] T. Syeda-Mahmood e D. Petkovic, "On describing color and shape information in images," *Signal Processing: Image Communication*, vol. 16, pp. 15-31, 2000.

[13] J. v. d. Weijer, C. Schmid e J. Verbeek, "Learning Color Names from Real-World Images," in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, 2007.

[14] S. Guberman, V. V. Maximov e A. Pashintsev, "Gestalt and Image Understanding," *Gestalt Theory*, vol. 34, nº 2, pp. 143-166, 2012.

[15] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh e R. v. d. Heydt, "A Century of Gestal Psychology in Visual Perception: I. Perceptual Grouping and Figure-Ground Organization," *Psychological Bulletin*, vol. 138, nº 6, pp. 1172-1217, 2012.

[16] J. Wagemans, J. Feldman, S. Gepshtein, R. Kimchi, J. R. Pomerantz, P. A. v. d. Helm e C. v. Leeuwen, "A Century of Gestalt Psychology in Visual Perception: II. Conceptual and Theoretical Foundations," *Psychological Bulletin*, vol. 138, nº 6, pp. 1218-1252, 2012.

[17] Stat Trek, "Linear Correlation Coefficient," 2012. [Online]. Available: http://stattrek.com/statistics/correlation.aspx. [Accessed on March 2nd, 2014].

[18] Barcelona Field Studies Centre, "Nearest Neighbour Analysis," May 11th 2013. [Online]. Available: http://geographyfieldwork.com/nearest_neighbour_analysis.htm. [Accessed on March 4th, 2014].

[19] N. Landsteiner, "meSpeak.js - Text-To-Speech on the Web," 2014. [Online]. Available: http://www.masswerk.at/mespeak/. [Accessed on April 11th, 2014]