DETECTION OF SEMANTIC ENTITIES USING DESCRIPTION GRAPHS

X.Giró-Nieto and F.Marqués

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya
{xgiro, ferran}@gps.tsc.upc.es

ABSTRACT

This paper presents a technique for the detection of Semantic Entities (SEs) in multimedia content. A definition of a SE in terms of lower-level SEs and their Relations (Rs) is proposed using Description Graphs (DGs). By analyzing the a/v information, an instance DG is built to be compared with a model DG of the SE. As a result, a confidence value is computed to express how well the SE is represented in the content. Examples of the use of this approach are presented in two different applications: detection of frontal faces and recognition of clusters of islands.

1. INTRODUCTION

The increase of digital multimedia content in the last years has created a need for an efficient indexing strategy in multimedia databases. A first approach to the problem is to describe content using low-level descriptors. In the case of still images and video sequences, an extensive work has already been produced in the definition and testing of visual descriptors [1]. For example, the MPEG-7 standard [2] for multimedia content description defines a set of visual descriptors for color, shape, texture and motion . However, accessing content through its low-level descriptors may not be flexible enough. Users commonly access databases searching for a Semantic Entity (SE) they have in mind. In many cases, this SE cannot be described only with a set of low-level descriptors, but it requires a structured combination of simpler SEs that present heterogeneous visual properties. This approach has opened the door to a new field of study aimed at finding semantic indexing solutions [3] [4]. This paper presents a technique for the semantic description of SEs using Description Graphs (DGs). The text is organized as follows. Section 2 provides a definition of DG and its different types of vertices. Section 3 presents a mathematical expression to evaluate the semantic similarity using DGs. Section 4 reports the application of DG techniques for candidate selection of human faces and clusters of islands. Finally, section 5 presents the conclusions of this work.

This work has been partly supported by a grant from of the Generalitat de Catalunya. This material is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA and the project IST-1999-20502 FAETHON

2. DEFINITIONS

2.1. Semantic Entities and Relations

A *Semantic Entity (SE)* is a generic term for referring to objects, agent objects, events, concepts, states, places, times and narrative worlds. A SE can be represented in several ways; e.g.: by a spatiotemporal segment in a multimedia document, by a low-level description as a set of visual descriptors, by a high-level description as a DG or by an alphanumeric label as a text annotation. For example, a football player can be represented by a region of pixels, by the color descriptors of his T-shirt, by a DG that consists of a head over a trunk and over two legs, or by the textual annotation used at the beginning of this sentence "football player". A *Relation (R)* is a feature that requires two or more SEs to achieve its full meaning. A Relation lacks any information when it is not related to any SE. The video sequence of a football goal event can be partitioned into two simple events like "Player hits the ball" and "Ball enters into the goal". These two SEs can be structured with temporal Rs like "Before" or "Later".

2.2. Description Graph

A generic graph is a group of V objects called *vertices* and a group of E non-ordered pairs of vertices called *edges*. A *Description Graph (DG)* assigns SEs and Rs to its vertices in order to describe a higher-level SE. SEs are linked by Rs using directed edges, in such a way that the neighborhood of the SEs are always Rs, and vice versa. At the same time, vertices are divided into necessary or optional. *Necessary* vertices must be present in any instance of a SE. On the other hand, the presence of *optional* vertices is not mandatory to have

an evaluable instance of a SE. Following the example of a goal event, the fact that the player celebrates the goal should be considered optional, because if he does not celebrate, it is still a valid goal. Nevertheless, the celebration reinforces the assumption that the previous scene represents a goal. As an example, a DG of the SE "Football goal" is shown in Figure 1.

Figure 1: *DG of the SE "Football goal".* Ellipses represent *SEs* and rectangles *Rs*. White denotes *necessary* and gray *optional*.

3. SE DETECTION

3.1. Models and instances

A SE detection algorithm can be designed based on DGs through a model/instance architecture. A *model DG* stored in an encyclopedia can be used as a reference for the detection of SEs in multimedia content. This *model DG* is compared to an *instance DG* generated as a result of an analysis of the content. A *confidence value C(SE)* expresses how well the SE is represented into the a/v document. In the case of a football goal, *C(Football goal)* will achieve its highest value when the analysis algorithm applied on a video sequence has an absolute certainty that the scene depicts a goal.

3.2. Vertex relevance and confidence

The computation of $C(SE)$ is performed based on the relevance and confidence of the vertices belonging to the model and instance DGs.

The *relevance (r)* of a vertex in a model DG expresses how important the presence of this vertex is for the complete description of the SE. Relevance takes values between 0 and 1, where 0 denotes irrelevant and 1 expresses absolutely relevant. In the previous example of a football goal, if the relevance of the vertex "Ball in goal" is higher than the one of vertex "Player hits the ball", a query for a goal to a database will retrieve before those sequences depicting only a ball in the goal than those showing only a player hitting a ball.

The *confidence (c)* of a vertex in an in an instance DG expresses how certain the analysis algorithm is of the presence of the SE or R in the instance. It is also defined in a scale between 0 (certainly no) and 1 (certainly yes). An automatic ball tracking algorithm could provide different degrees of confidence when detecting that the object "Ball" is in the object "Goal".

3.3. SE confidence measure using DG

The comparison of model and instance DGs is calculated by applying Equation 1. Relevances of model vertices (r_i) are combined with confidences of instance vertices (c_i) to obtain a global $C(SE)$ for the whole structure. The expression is normalized by the relevances to obtain values between 0 and 1. Sums of the N necessary vertices and the O optional vertices are expressed separately for clarity.

$$C(SE) = \frac{\sum_{i=0}^{N} r_i c_i + \sum_{i=0}^{O} r_i c_i}{\sum_{i=0}^{N} r_i + \sum_{i=0}^{O} r_i} \quad (1)$$

When computing $C(SE)$, all necessary vertices must be included. On the other hand, there is no information a priori to decide which optional vertices must be in the expression. A selection criterion must be defined to choose only those optional vertices that will increase $C(SE)$. As shown in Equation (2), only those optionals vertices with a confidence higher than the current $C(SE)$ must be included in the expression. With this criterion, the design of the decision algorithm is simple. All optional vertices are firstly sorted in decreasing order according to their associated c. The ordered list is scanned adding optional vertices to the expression until reaching a vertex with a c smaller than the current C(SE).

$$\frac{\sum_{i=0}^{N} r_i c_i + \sum_{i=0}^{j-1} r_i c_i + r_j c_j}{\sum_{i=0}^{N} r_i + \sum_{i=0}^{j-1} r_i + r_j} > \frac{\sum_{i=0}^{N} r_i c_i + \sum_{i=0}^{j-1} r_i c_i}{\sum_{i=0}^{N} r_i + \sum_{i=0}^{j-1} r_i} \Leftrightarrow$$

$$\Leftrightarrow c_j > \frac{\sum_{i=0}^{N} r_i c_i + \sum_{i=0}^{j-1} r_i c_i}{\sum_{i=0}^{N} r_i + \sum_{i=0}^{j-1} r_i} \quad (2)$$

3.4. Simple SE confidence measure using low-level descriptors

The similarity between an instance and a model is not only restricted to DGs. For example, the simplest SEs are not represented by any DG but by their low-level descriptors. In this work, MPEG-7 visual descriptor have been used. In the case of football, the ball can

be defined as a round and white object. For simple SEs, a low-level relevance and confidence metric must be defined for each descriptor. These relevances and confidences are also combined using Equation 1.

4. STUDY CASES

DGs have been tested in the selection of the image candidates to best represent two completely different SE, "Frontal faces" and "Group of islands".

4.1. Frontal faces

The output candidates generated by a frontal face detection algorithm [5] were evaluated using DGs techniques. From an initial generic partition, the analysis algorithm extracted several candidates to be a frontal face, as well as some facial features with their associated confidences. An instance DG of each of the candidates was generated based on the model DG shown in *Figure 2*. This model considers that facial features like eyebrows or nostrils are optional, because in many cases they may be occluded by hair or hidden due to light effects. The candidate with the highest $C(SE)$ was selected as the one best representing a frontal face, as shown in *Figure 3*. Experiments showed that the flexibility introduced by the optional vertices helped the algorithm to select the best candidate.

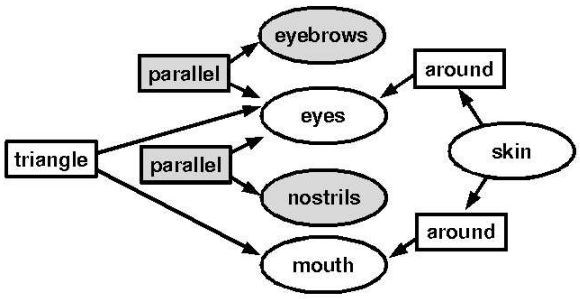

Figure 2: *Model DG of the SE "Frontal face".*

4.2. Clusters of islands

The flexibility of DGs is further demonstrated on the detection of clusters of islands. The DG shown in Figure 4 represents a cluster of four islands in the archipelago of the seven Canary Islands. The model DG shown in Figure 4 was built by creating a SE for each island. For each of them, the MPEG-7 contour-shape visual descriptor [6] was extracted. The four triangles defined by the center of masses of the islands defined a Relation. Triangular shapes were chosen given their

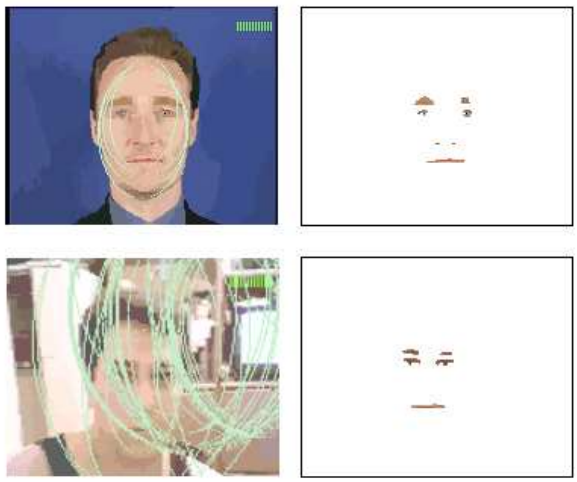

Figure 3: *Results of detection of SE "Frontal face". Candidates are represented by ellipses.*

invariance to rotation and scaling [7]. The model was compared with a set of test images of the area that had been segmented using the technique presented in [8]. The experiment consisted in testing all possible combinations resulting from identifying each of the seven segments to every of the four SEs. In every case, the $C(SE)$ was computed and the highest score was selected as the best representation of the group of four islands. The generated results, exemplified in Figure 5, showed how the triangular relation helped to overcome the contour deformation produced by cloud occlusion and the geometrical distortion that appears in all satellite images.

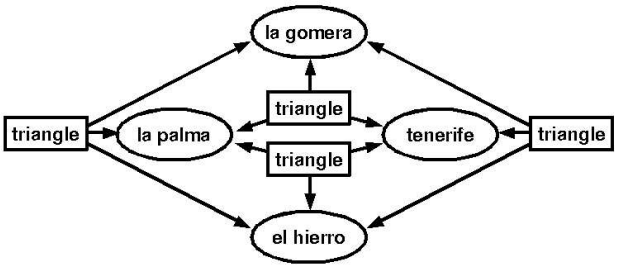

Figure 4: *Model DG of SE "Group of islands".*

5. CONCLUSIONS

Description Graphs provide a generic technique to evaluate the presence of SEs in multimedia content. The

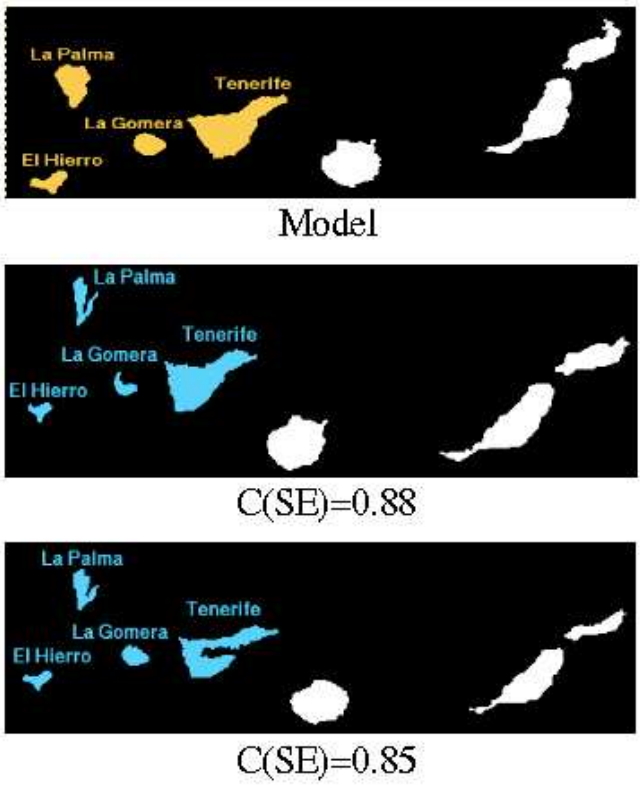

Figure 5: *Model and candidate images of SE "Groups of islands".* Grey areas correspond to model and selected islands.

structured description of a SE in terms of lower level SEs and their Rs allow their detection even when they are defined by non-homogeneous visual descriptors. The model/instance architecture, combined with the predefined relevances and the extracted confidences, allows the computation of a confidence value that expresses how well a SE is represented in an a/v document.

6. REFERENCES

[1] M.A.Nascimiento, V.Sridhar and X.Li, *Effective and efficient region-based image retrieval.* Journal of Visual Languages and Computing 2000

[2] B.S.Manjunath, P.Salembier and T.S.Sikora. *Introduction to MPEG-7.* Wiley 2002

[3] A.B.Benítez and S.Chang. *Semantic knowledge construction from annotated image collections.* IEEE International Conference on Multimedia and Expo (ICME), Switzerland, 2002

[4] A.Ekin, A.Murat Tekalp and R.Mehrotra. *Integrated semantic-syntactic video event modeling for search and retrieval.* IEEE International Conference on Image Processing (ICIP), Rochester, New York, United States, 2002

[5] F.Marqués and C.Sobrevals. *Facial Feature Segmentation from Frontal View Images.* Proceedings of the Eleventh European Signal Processing Conference EUSIPCO-2002, pp, 33-36 Tolouse, France, September 2002

[6] M.Bober. *MPEG-7 Visual Shape Descriptors.* IEEE Transactions on Circuits and Systems for Technology, Vol.11, No.6, June 2001

[7] M.Isaksson. *Face detection and pose estimation using triplet invariants.* www.ep.liu.se/exjobb/isy/2002/3223/exjobb.pdf

[8] F.Eugenio, F.Marqués and J.Marcelo. *Pixel and sub-pixel accuracy in satellite image referencing using an automatic contour matching approach.* Proceedings of the IEEE International Conference on Image Processing ICIP-2001, pp. I.822 - I.825, Thessaloniki, Greece, October 2001.