Escuela Politécnica Superior
Universidad Autónoma de Madrid

“Extraction of mid/high-level semantic features for the indexation and classification of television journals videos”

(Extracción de características de medio y alto nivel semántico para la anotación y clasificación de videos de noticias de televisión)

Javier Molina Vela

Tutor: José María Martínez

Abbreviations and Acronyms:

MPEG	Moving Pictures Expert Group
CLD	Color Layout Descriptor
contour-SD	Contour Shape Descriptor
CSD	Color Structure Descriptor
DCD	Dominant Color Descriptor
HTD	Homogenous Texture Descriptor
region-based SD	region-based Shape Descriptor
SCD	Scalable Color Descriptor
SVM	Support Vector Machine
TV	Television
VC	Visual Cue
XM	eXperimentation Module

Table of Contents

1. Introduction

1.1. Motivation

During the last decades, a huge amount of multimedia information has been (and is being) generated in many different contexts (i.e. movies, serials, musical videos, news…). Among these categories of content, news production deserves special attention, since it is one of the environments in which video sequence retrieval is more frequently needed and moreover new content is generated at a daily basis. It is very common in television (TV) journals, to accompany a news report with archived material of similar news, besides the up-to date contributed footage. Thus, the need for annotation of the huge audiovisual information appears necessary. This annotation cannot be achieved by simply associating keywords manually to each video for two main reasons. The first one is the exponentially increasing quantity of digital content, which would make it a tedious and time-consuming task. The second is that the content of a multimedia document cannot be fully described by a simple list of words but in terms of a visual representation. Thus, the extraction of visual information directly from the multimedia documents is required. This procedure is usually called "low-level feature extraction". Many different techniques have been developed for describing local image regions in terms of their features. These features can be efficiently captured by standardized visual descriptors, such as MPEG-7 descriptors [1][2]. Although there exists a great variety of visual descriptors, standardized or not, the choice of the appropriates ones is not a trivial task, since one should consider many different parameters such as the discrimination capabilities, the time required for their extraction, their levels of detail for scalability support, etc.

The Mesh (*Multimedia Semantic Syndication for Enhanced News Services)* project faces, among others, the topics enumerated above. It is under the European Community's Sixth Framework Programme and the research leading to this document has received funding from this programme.

1.2. Objectives

The main targets of this work are:

- To evaluate the State of the Art and existing background technologies and its suitability to obtaining the first estimations of classification accuracy and computational cost of most widely used visual descriptors.
- One of the objectives of this research work, under the domain of video news reports, focuses on the selection of the most appropriate MPEG-7 visual descriptors and their corresponding levels of detail for achieving a certain balance between computational cost and classification accuracy of spatial regions. More specifically, in this paper we propose a method for studying, in the specific context of nature disasters, the relation between the classification accuracy of spatial regions and the computational cost of the required extractions of descriptors along with their levels of detail. This information will be useful for later implementations of real-time classifiers of video shots based on descriptor extraction in spatial regions. An important aim of this work is to obtain a methodology applicable to different contexts (i.e. sports, weather reports…).
- To compare data clustering accuracies with the ones obtained by means of a Support Vector Machine (SVM), in order to evaluate the reliability of these last ones.

- To present the current research line which faces image level annotated data sets, rather than manually segmented and annotated spatial regions. Due to the huge amount of elements to be clustered, a method based on a semantically guided data clustering is proposed.
- To present preliminary results and future required work.

1.3. Structure of the Report

The remaining of the report is structured as follows: Section 2 presents the State of Art on the two topics of this work: visual descriptors and separation of objects and region recognition techniques. In section 3 the content sets over which the experiments have been performed, is described. Section 4 focuses on the description of the MPEG-7 visual descriptors suitable for this work, as well as the levels of detail. Some descriptors are also ruled out because of their context of application (which is not spatial regions) or their documented too high extraction time. In section 5 the proposed method for spatial regions separation is described, the resulting relations between classification accuracy of spatial regions and computational cost for some interesting profiles are presented and the particularities commented. A comparison with data clustering techniques is presented in order to estimate the reliability of the results.In section 6 some conclusions and current and future lines of action are enumerated.

In A.1 the whole collection of classification accuracies and computational costs estimated in section 5 is presented. In A.2 an approach to the problem of working with globally annotated images is presented.

Some of the contents presented in sections 3.1, 4 and 5 conform a paper entitled “On the selection of MPEG-7 Visual Descriptors and their Level of Detail for Nature Disaster Video Sequences Classification” that has been accepted for oral presentation at SAMT 2007 (Semantic and Digital Media Technologies 5-7 December 2007, Genova, Italy) [**http://samt2007.ge.imati.cnr.it/**] and will be published in the conference Proceedings (LNCS).

2. State of Art

2.1. Visual Descriptors

Performance evaluation has gained more and more importance in multimedia analysis applications, as a mandatory step for implementation of real-time systems. Very little work has been done on the evaluation of efficiency extraction of local descriptors in the context of matching and recognition and classification [3]. The main focus of the performance evaluation task is that the descriptors should be distinctive and at the same time robust to changes in viewing conditions as well as errors of the detector. Several studies about MPEG-7 descriptors and their accuracy in image retrieval have been performed [4][5][6][7][8].

2.2. Object/Region Categorization Techniques

Various approaches can be found to the problem of objects/regions categorization. Two main difficulties are found in the task of labelling spatial regions with high/mid level semantic concepts. The first is the absence of a reliable and generic spatial segmentation method. The second is the limitations of the low-level features for separating higher level concepts, that is, to bridge the "semantic gap".

An important opened line of research to solve the segmentation problem consists on performing image segmentation and automatic semantic annotation at same time [9] [10]. This way the use of semantic information improves the spatial segmentation and *vice versa*. This methodology is useful when detecting presences which usually present homogenous in colour and/or in texture spatial regions (*i.e.* water, sky, smoke…). In [10] a solution based on spatial regions clustering is proposed for image separation in semantic clusters.

Other approaches focus on the use of visual words for recognizing certain objects [11][12][13]. These methods work when having previously a reliable segmentation algorithm or when having images where the objects to be recognized occupy the whole frame.

In [14] a system based on low-level image regions clustering is proposed, applying at the same time semantic pair-wise constraints (i.e. regions coming from images with sky, water and vegetation cannot be merged with regions that belong to indoor images). This points out an interesting line for future improvements of our system. In [15] another example in which spatial regions are the elements to be clustered using low level features can be found.

3. Content Set used as Ground-truth

As usual in image classification problems, the ground truth is one of the most important restrictions for obtaining reliable results, as it will be used for training the classification machines. It has to contain a significant sample of each of the possible classes under study.

Since the aim of this study is to compile useful knowledge for later use in TV journals real time automatic annotation, the content set should contain visual cues usually present in TV journals.

Two different content sets are presented in this section: a spatial image region manually segmented and annotated (see 3.1) and a larger image collection manually annotated without considering spatial regions (see 3.2). The first of them is used for estimating the relation between classification accuracy and cost estimation (see 5.1). The second is nearer to the commonly found ground-truths in image or video semantic information retrieval approaches. This is why, although no conclusive methodology is presented for this kind of image sets, it is of interest to introduce at least one of them.

3.1. Region Level Ground-truth

Due to the enormous human cost necessary for the compilation of a manually segmented and annotated collection, the semantic concepts taken into account need to be reduced. In this case, to the context of nature disasters news (see 3.1.1).

3.1.1. Target Classes

Although the aim of this paper is not the detection of semantic cues, its enumeration allows the definition of which visual cues are useful to detect. The context will be nature disasters TV journals (see Table 1).

Table 1: Visual Cues selected based on nature disaster semantic cues.

Semantic Cues	Visual Cues
Forest fires	flames, smoke, vegetation
Urban fires	flames, smoke, buildings
Floods/Tsunamis	water
Hurricanes/typhoons/windstorms	water
Earthquakes	rocks, buildings
Volcanic eruptions	flames, rocks
Snow avalanches	snow

So, the target classes (visual cues) in this descriptors' performance study will be: water, flames, smokes, vegetation, rocks, buildings and snow.

3.1.2. Data Set

The dataset is composed of a subset of MESH video repository, images from the Labelme dataset [16] and various images collected from the *world wide web*. The manually segmented regions used in the training process have been obtained using *Labelme annotation tool* [16]. Around one hundred spatial regions of each visual cue have been compiled. Some examples are shown in Figure 1.

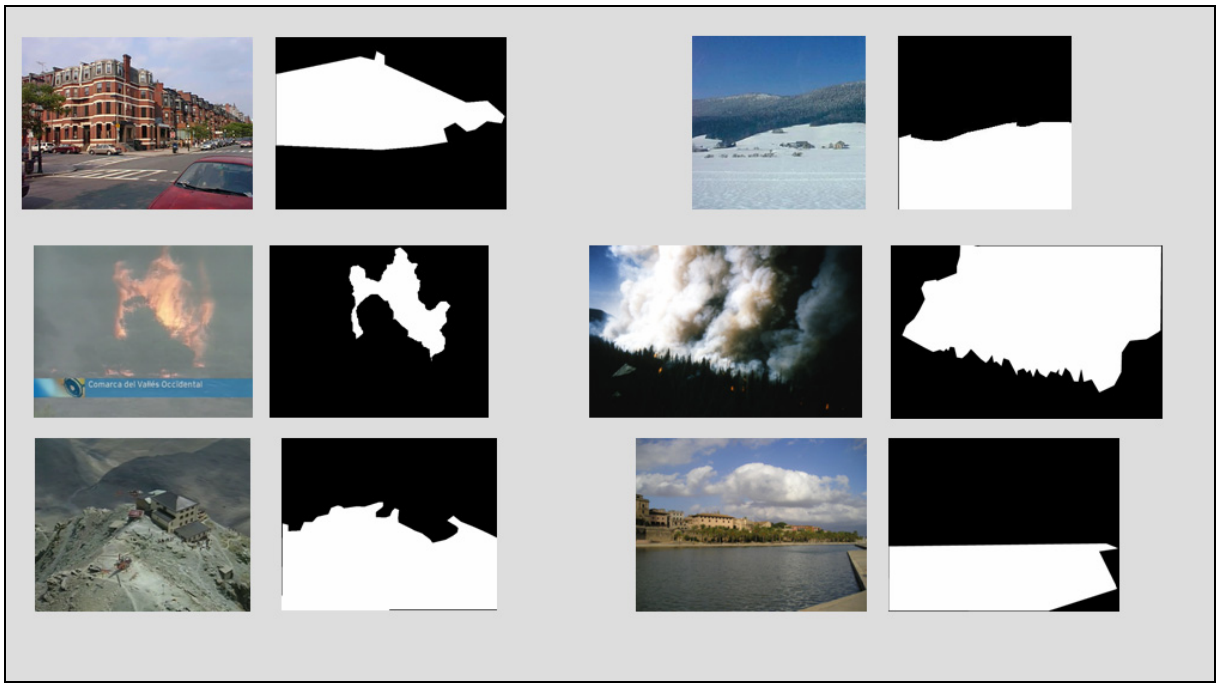

Fig. 1. Examples of manually segmented spatial regions.

3.2. Image Level Ground-truth

The most commonly accessible ground-truth in video and image semantic information retrieval studies consist of images annotated with certain mid/high-level semantic concepts. This is, with no manual spatial segmentation, which has an enormous human cost. A first approach to solve the associated problematic is described in A.2.

3.2.1. Target Classes

The target classes or visual cues considered on this study are all closely related with TV news journals. Unlike to the regions ground truth described in 3.1, the visual cues chosen for being annotated are not restricted to nature disasters. In this case the semantic concepts are related with TV journals in general.

The target classes are the ones defined in the TRECVID 2007 guidelines [17], and they are the following:

1. Sports: Shots depicting any sport in action
2. Entertainment: DROPPED
3. Weather: Shots depicting any weather related news or bulletin
4. Court: Shots of the interior of a court-room location
5. Office: Shots of the interior of an office setting
6. Meeting: Shots of a Meeting taking place indoors
7. Studio: Shots of the studio setting including anchors, interviews and all events that happen in a news room
8. Outdoor: Shots of Outdoor locations
9. Building: Shots of an exterior of a building
10. Desert: Shots with the desert in the background

11. Vegetation: Shots depicting natural or artificial greenery, vegetation woods, etc.
12. Mountain: Shots depicting a mountain or mountain range with the slopes visible
13. Road: Shots depicting a road
14. Sky: Shots depicting sky
15. Snow: Shots depicting snow
16. Urban: Shots depicting an urban or suburban setting
17. Waterscape_Waterfront: Shots depicting a waterscape or waterfront
18. Crowd: Shots depicting a crowd
19. Face: Shots depicting a face
20. Person: Shots depicting a person (the face may or may not be visible)
21. Government-Leader: DROPPED
22. Corporate-Leader: DROPPED
23. Police_Security: Shots depicting law enforcement or private security agency personnel
24. Military: Shots depicting the military personnel
25. Prisoner: Shots depicting a captive person, e.g., imprisoned, behind bars, in jail or in handcuffs, etc.
26. Animal: Shots depicting an animal, not counting a human as an animal
27. Computer_TV-screen:Shots depicting a television or computer screen
28. Flag-US: Shots depicting a US flag
29. Airplane: Shots of an airplane
30. Car: Shots of a car
31. Bus: Shots of a bus
32. Truck: Shots of a truck
33. Boat_Ship: Shots of a boat or ship
34. Walking_Running: Shots depicting a person walking or running
35. People-Marching: Shots depicting many people marching as in a parade or a protest
36. Explosion_Fire: Shots of an explosion or a fire
37. Natural-Disaster: Shots depicting the happening or aftermath of a natural disaster such as earthquake, flood, hurricane, tornado, tsunami
38. Maps: Shots depicting regional territory graphically as a geographical or political map
39. Charts: Shots depicting any graphics that is artificially generated such as bar graphs, line charts, etc. (maps should not be included)

Some of the originally proposed visual cues are *dropped* for the unfeasibility of being detected or identified or simply because a lack of accessible contents.

For the interest of this study, we will distinguish between to groups of visual cues:

- The presences - Understood as visual cues which are defined by the presence or absence of a certain physic reality, identifiable in a spatial region of the image. These are: building, vegetation, mountain, road, sky, snow, waterscape_waterfront, crowd, face, police_security, military, prisoner, animal, computer_TV-screen, Flag-US, airplane,

car, bus, truck, boat_ship, walking_running, people_marching, explosion_fire, maps, charts.

- The concepts - Understood as visual cues which are defined representation of concept non identifiable in a concrete spatial region. They are the following: sports, weather, court, office, meeting, studio, outdoor, desert, urban, natural-disaster.

3.2.2. Data Set

The set of images used as ground-truth is the one provided the 2007's edition of the TRECVID contest [17]. This collection consist of around twenty thousand manually annotated images, for what a collaborative annotation tool [18][19] was developed *ex profeso*. As result of this annotation process we got four possible annotations values for each visual cue presence/incidence: *positive*, *negative*, *skipped* or *no annotated*. During the manual annotation process, images are marked as *skipped* when the annotator is not sure of the presence/absence or incidence/"no incidence" of the visual cue.

4. Visual Descriptors Selection

4.1. Overview of Visual Descriptors

The MPEG-7 eXperimentation Model [20] has been designed to allow the extraction of all visual and multimedia descriptions[1] illustrated in the standard [2]. Moreover, it allows, for certain descriptors, the extraction with various levels of detail: The *Dominant Color Descriptor*, the *Scalable Color Descriptor*, the *Color Structure Descriptor* (see 4.1.1); the *Homogeneous Texture Descriptor* (see 4.1.2) and the *Texture Browsing Descriptor*. As well, some of the proposed descriptors have only one working level of detail: the *Color Layout Descriptor* (see 4.1.1), the *Region-Based Shape Descriptor* and the *Contour-Shape Descriptor* (see 4.1.3). In this section, different levels of detail are proposed for studying the relation between their computational cost and the resulting accuracy in classification results.

Since this work gives focus to spatial regions of still images, it appears that some visual descriptors do not make sense in the presented application field. First of all, motion descriptors are discarded, because they all need the temporal coordinate to be calculated and currently we are working at the frame-by-frame level. One of the main future research lines is focused on the analysis of shape evolution in spatiotemporal regions. The *Group-of-Frame/Group-of-Picture Descriptor* is discarded since it is used for joint representation of a group of frames or pictures. Moreover, the *3-D Shape Descriptor* is directly ruled out from this study, simply because we are working with 2-D projections of the 3-D real world and, since our first approach is based on the study of still images, it is not possible to create 3-D regions as a merging of 2-D spatial regions obtained from several views of the same object. The *Edge Histogram Descriptor* is only applicable to rectangular images, no to arbitrary shape regions, and its use makes sense for global extraction from an image (no from a region).

The final aim of this study is to compile useful information for later implementations of real-time visual descriptions extraction systems for video sequences and the classification of the regions of the images based on those descriptions. Towards this goal, it makes sense to rule out the *Texture Browsing Descriptor*, since its extraction (with the XM) is about 20 times slower than the *Homogenous Texture Descriptor* as documented in [8]. This, in combination with the measurements presented in section 5.1, implies that the extraction of this descriptor is 28 and 19 times slower than the slowest extractions of colour and shape descriptors.

In the following subsections an overview of the descriptors under study and their available levels of detail is presented.

4.1.1. Colour Descriptors

Since colour is probably the most discriminative of all visual features, the MPEG-7 standard offers a variety of colour descriptors [22]. In brief, these descriptors consist of colour histograms, a *Dominant Color Descriptor* and a *Color Layout Descriptor* and have been designed in order to be applied in a wide range of applications:

- The *Dominant Color Descriptor* (DCD) provides a compact and not fixed to a specific colour-space description of an image or an image region. After a clustering of the present colours within the image/region, up to 8 representative colours are extracted, along with their percentages and variances. This way it offers compact and intuitive information of the present colours within the image/region. It is possible to select the number of bins (NB) for each colour channel, among eight different proposed: 256,

[1] Visual description refers to the specific values of visual descriptors associated to an element.

128, 64, 32, 16, 8, 4 and 2. MPEG-7 proposes three ways of similarity matching. The first one uses the extracted colours and their percentages, the second considers also the spatial coherency (SP) and the third uses the variances (variances present, VP). This makes 32 different levels of detail.

- The *Scalable Color Descriptor* (SCD) provides a colour histogram in the HSV colour space, encoded with a Haar transform–based scheme. It is more verbose than *DCD* and designed to allow scalability. It has two configurable parameters: the number of bit planes discarded (NBPD) (0, 1, 2, 3, 4, 5, 6 or 8) and the number of coefficients (NC) (16, 32, 64, 128 or 256). This makes 40 different levels of detail.
- The *Color Structure Descriptor* (CSD) is the most complete image/region level colour descriptor. It contains information both of the colour distribution of an image in a similar way to a colour histogram and of the local spatial structure of the colour. This descriptor is formed as a histogram and the colour quantification resolution (QR) can be set up with the following values: 256, 128, 64 and 32, resulting in 4 different levels of detail.
- The *Color Layout Descriptor* (CLD) offers a compact and resolution-invariant representation of the spatial distribution of colours for high-speed image retrieval. The same information could be obtained by the use of the *Dominant Color Descriptor* applied to a grid of the image, but, this way, the number of bits is greater, complicating its use as a pattern for a classification machine. It has no configurable parameter, so it has only one level of detail.

4.1.2. Texture Descriptors

Another very important visual feature of images is texture. Many natural and artificial objects can effectively be described by their texture and many applications such as aerial imagery have successfully used texture descriptors. In this study, we focus on one of the MPEG-7 texture descriptors [22]:

- The *Homogeneous Texture Descriptor* (HTD) shows information of the texture by means of the mean energy and the variance of an image or region, for 30 different 2-D frequency channels which are modelled using *Gabor* functions. The 2 levels of detail will be defined by the calculation or not of the energy variance for each of the channels (layer 1 and layer 0).

4.1.3. Shape Descriptors

For certain search and retrieval applications, the shape of objects appears a very powerful visual characteristic. Here we use two of the MPEG-7 shape descriptors [23]:

- The *Region-Based Shape Descriptor* (region-based SD) is a shape descriptor able to represent, in a compact fashion, multiple regions objects, or regions with holes. It has only one working level of detail.
- The *Contour-Shape Descriptor* (contour-SD), contrarily to the region-based one, focuses in the differences between the contours for increasing its discrimination capacity between regions. It has only one working level of detail.

4.2. Dissimilarity Measures for Visual Descriptors

In [2], distances for each visual descriptor are proposed. Here they are briefly described:

- For the *Dominant Color Descriptor* (DCD):

$$D^2(F_1,F_2)=\sum_{i=1}^{N_1}p_{1i}^2+\sum_{j=1}^{N_2}p_{2i}^2-\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}2a_{1i,2j}p_{1i}p_{2j}$$

where:

- $F_1=\{\{c_{1i},p_{1i},v_{1i}\},s_1\},i=1,2,...,N_1$ and $F_2=\{\{c_{2i},p_{2i},v_{2i}\},s_2\},i=1,2,...,N_2$, are two dominant colour descriptors and N_1 may be different than N_2.
- $a_{k,l}$ is the similarity coefficient between two colours c_k and c_l, defined by: $a_{k,l}=\begin{cases}1-d_{k,l}/d_{\max}, d_{k,l}\le T_d\\ 0, d_{k,l}>T_d\end{cases}$ and $d_{k,l}$ is the Euclidean distance between the two colours c_k and c_l. Moreover, T_d is the maximum distance for two colours and $d_{\max}=aT_d$.

For a more detailed description of this similarity measure, the reader may refer to the MPEG-7 eXperimentation Model (XM) [20], where some details for the selection of both T_d and a can be found.

- For the *Scalable Color Descriptor* (SCD): MPEG-7 does not strictly define a similarity measure for this descriptor. In similarity matching of histograms, the L1 norm (sum of absolute differences) usually results in good retrieval accuracy [2].
- For the *Color Structure Descriptor* (CSD): Since this descriptor has the form of a histogram and MPEG-7 does not define its own similarity measure, all popular and applicable to histogram measures such as *Hamming* or *Euclidean* distance can be used.
- For the *Color Layout Descriptor* (CLD): The similarity measure applied on the CLD descriptor is the following:

  $$D=\sqrt{\sum_i w_{yi}(DY_i-DY_i')^2+\sum_i w_{bi}(DCb_i-DCb_i')^2\sum_i w_{ri}(DCr_i-DCr_i')^2}$$

  where DY_i, DCb_i and DCr_i represent the *i*-th DCT coefficients of the respective colour components.
- For the *Homogeneous Texture Descriptor* (HTD): The similarity measure of the HTD is the following: $d(TD_{query},TD_{database})=\sum_k\left|\frac{TD_{query}(k)-TD_{database}(k)}{a(k)}\right|$. Here, the absolute difference between two sets of feature vectors is considered, one for the query image TD_{query} and one for the image in the database $TD_{database}$. The recommended normalization value $a(k)$ is the standard deviation of $TD_{database}(k)$. The user can specify its own choice for $a(k)$, for example, $a(k)=1$, which would make this distance the Hamming distance.
- For the *Region-Based Shape Descriptor* (region-based SD): The dissimilarity measure proposed in [2] is the L-1 norm (the *Hamming* distance).
- For the *Contour-Shape Descriptor* (contour-SD): A more complex than the *Hamming* or *Euclidean* dissimilarity measure is proposed for this descriptor. It is based on the number and characteristics of the peaks on the contour, as well as its eccentrity and circularity. For a more detailed explanation see [2].

5. Spatial Region Separation

In the research work presented here, we propose a methodology that, operating on local feature extraction (image regions), has as aim to finally perform a separation at an image level. The use of spatial regions of interest instead of whole images in the low-level features extraction allows the acquisition of more reliable values for the colour and texture descriptors while opening the way to the use of shape descriptors.

In this study a SVM is used (see section 5.2), this implies that the dissimilarity measure (or distance) between the elements is the *Euclidean distance*. As commented in 4.2, the *Euclidean distance* is applicable to some of the descriptors, but for some other descriptors more complex distances are proposed. We assume that, although the classification results for the profiles and levels of detail analyzed in some cases are not the best, the results will point out which combinations of descriptors and levels of detail should be taken into account for a deeper study with data clustering techniques (see section 5.3) in order to quantify their reliability. SVM presents certain advantages in comparison with data clustering techniques. In test or validation, the categorization consists on a fast identification of the group to which the element belongs, while with data clustering techniques the distances from the element to each of the clusters has to be calculated. In some contexts the distances are to heavy and the number of clusters too high for achieving a real-time categorization.

5.1. Extraction Computational Cost Estimation

Within the extraction of the aforementioned visual descriptors (see 4.1), calculations like colour clustering or quantification, which make hard an analytical estimation of the computational costs, usually take place. Moreover, the time invested in them depends on the specific images being analyzed.

In this section, we present the results of measuring the extraction time of the descriptors and their levels of detail, proposed in section 4.1, over the content set described in detail in section 3. The extraction times have been measured while executing the XM version 6.1 [20] on an Intel(R) Core(TM) 2 Duo CPU T7200 @ 2Ghz with 1GB of RAM. The time invested by the application for loading images and masks in memory has been subtracted.

The presentation of the results is done by joining different levels of detail in each descriptor when they show low variance in their computational cost (see Table 1). The extraction time is the mean for each descriptor. Its percentage variation, when present, expresses the deviation between the grouped, because of their low computational cost differences, levels of detail. The different labels used in the first and second column correspond to the ones suggested in section 4.1.

Table 1: Levels of detail Extraction times.

Label	Descriptors and levels of detail	Extraction time (msec)
DCD1	DCD: VP (1,0), BN (256,128,64,32,16,8,4,2), **SC (1)**	$1832 \pm 0.50\%$
DCD2	DCD: VP (1,0), BN (256,128,64,32,16,8,4,2), **SC (0)**	$1577 \pm 1.10\%$
SCD	SCD: NBPD(0,1,2,3,4,5,6,8), NC(256,128,64,32,16)	$196 \pm 0.05\%$
CSD	CSD: QR(256,128,64,32)	$191 \pm 0.15\%$
CLD	CLD	65
HTD	HTD: layer 1 or layer 0	$2652 \pm 0.05\%$
rSD	region-based SD	1933

cSD	contour-SD	315

5.2. Classification Accuracy Estimation

To be rigorous, the classification accuracy would need to be calculated for all the possible combinations of descriptors and levels of detail. But, as shown in section 5.1 many levels within each descriptor present similar computational costs in their extraction. So, in this context, it makes sense to use the most detailed level of the ones belonging to the group of similar computational cost in order to represent all the rest as one of the ultimate goals is also classification accuracy.

To take into account the possibility of not using all the descriptors in all the combinations, a new level for each descriptor is introduced, one that represents the non usage of the corresponding descriptor. Using the equation (1) with the values compiled in Table 2, the obtained number of patterns´ sets (Ns) is 96.

$$Ns = np_{DCD} \times np_{SCD} \times np_{CSD} \times np_{CLD} \times np_{HTD} \times np_{\text{region-based SD}} \times np_{\text{contour-SD}} \quad (1)$$

Table 2: Number of levels of detail per descriptor (the final +1 means no appearance of that descriptor)

Descriptor	Number of levels of detail
Dominant Color Descriptor	$np_{DCD} = 2 + 1$
Scalable Color Descriptor	$np_{SCD} = 1 + 1$
Color Structure Descriptor	$np_{CSD} = 1 + 1$
Color Layout Descriptor	$np_{CLD} = 1 + 1$
Homogenous Texture Descriptor	$np_{HTD} = 1 + 1$
Region-Based Shape Descriptor	$np_{\text{region-based SD}} = 1 + 1$
Contour-Shape Descriptor	$np_{\text{contour-SD}} = 1 + 1$

Each of these pattern sets is analyzed with a SVM using the LIBSVM implementation [24] , making 10 folds and calculating the classification accuracy for the 10 combinations 90% training/10% testing with RBF kernel. The resulting separation accuracy is the mean of all those. The relations of computational cost and classification accuracy for each of these levels of detail are presented in section 5.4.

5.3. Spatial Regions Clustering

As already commented, the SVM does not make use of the proposed distances for each descriptor (see section 4.2). In order to measure the reliability of the classification accuracy obtained with this technique, a more rigorous approach based on data clustering is proposed.

The approach consists of a hierarchical clustering [25] in which the stopping criterion is based on the inconsistence of the clusters. The separation accuracy for each concept is estimated as the percentage of elements of that concept that constitute a simple majority in any of the clusters. The simple majority of a cluster implies the assignment of the label of this

majority to the cluster. In Fig. 2 we find an extract of the incidences of the collection of clusters. The numbers which are simple majority are written in bold.

[CSD]	VC1	VC2	VC3	VC4	VC5	VC6	VC7
Cluster 1	0	**2**	0	0	0	0	0
Cluster 2	0	**6**	0	0	0	0	0
Cluster 3	0	**4**	0	1	0	0	0
Cluster 4	0	**3**	0	0	0	0	0
Cluster 5	**2**	1	0	0	0	0	0
Cluster 6	0	**2**	0	2	0	0	0
...	...	...	...	...	...	...	...

VC1: buildings VC2: fire VC3: rocks VC4: smoke
VC5: snow VC6: trees VC7: water

Fig. 2. Extract of incidences of clusters.

5.4. Results Analysis

5.4.1. SVM separation

In this section, some particularities of the relation between classification accuracy and computational cost for the different profiles are commented. The complete compilation of these values is exposed in detail in appendix A.1.

In Figure 3, each point represents each of the 96 possible profiles. The asterisks represent the best profiles (the ones that manage the highest classification accuracies, extracted from A.1) within a computational cost interval (see Table 3). As we can see, in some cases, similar accuracy can be managed with different computational costs. The computational costs are expressed as the percentage of the maximum value of computational cost (7184 msec.) which corresponds to the profile [DC1 SCD CSD CLD HTD rSD cSD] , which implies the extraction of all the descriptors and for the *Dominant Color Descriptor* the most detailed level (see 5.1).

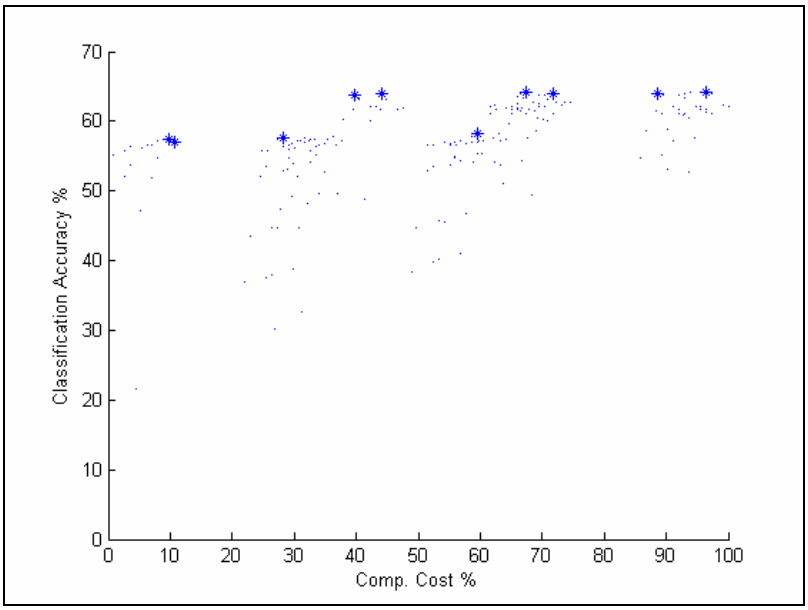

Fig. 3. Relation between computational cost and classification accuracy for all the possible combinations of descriptors and levels of detail.

Table 3: Relations between computational cost and classification accuracy for the best profiles of computational cost segments.

Profile	Cost (%)	Acc. (%)	Acc. /Cost
[SCD CSD cSD]	(0,10]	57.41	5.8762
[SCD CSD CLD cSD]	(10,20]	56.98	5.3352
[DC2 SCD CSD CLD]	(20,30]	57.56	2.0382
[SCD HTD]	(30,40]	63.66	1.606
[SCD HTD cSD]	(40,50]	63.95	1.4524
[DC2 SCD CSD CLD HTD]	(50,60]	61.63	0.94583
[SCD CLD HTD rSD]	(60,70]	64.24	0.95227
[SCD CLD HTD rSD cSD]	(70,80]	63.95	0.89017
[DC2 SCD HTD rSD]	(80,90]	63.95	0.7226
[DC1 SCD HTD rSD cSD]	(90,100]	64.1	0.66466

It can be inferred from the compiled data in Table 3 and Figure 3 that, in some combinations, the use of a specific descriptor is counterproductive, making worse the classification results, as well as, of course, increasing the extraction time. It is not always accomplished that the more computational cost it is invested the better classification accuracies are obtained (i.e. the highest classification accuracy requires a computational cost between 60 and 70% of the maximum).

As far as we are focusing this study on obtaining useful information for later real time implementations, it makes sense to highlight the profiles which show highest accuracy/computational cost ratios (see Table 4), even when the classification results are not the best ones. This gives us a vision of which MPEG-7 visual descriptors are most recommendable for real time contexts, coming up the *Color Layout Descriptor* as the most efficient.

Table 4: Profiles with best relation between accuracy and computational cost.

Profile	Cost (%)	Acc. (%)	Acc. /Cost
[CLD]	0.9	55.09	61.211
[SCD]	2.73	55.81	20.443
[CSD]	2.66	52.18	19.617
[SCD CLD]	3.63	56.4	15.537
[CSD CLD]	3.56	53.78	15.107
[SCD CSD]	5.39	56.25	10.436
[SCD CSD CLD]	6.29	56.69	9.0127
[CLD cSD]	5.29	47.09	8.9017
[SCD cSD]	7.11	56.69	7.9733
[CSD cSD]	7.04	51.89	7.3707

5.4.2. Data Clustering Separation

In this section, the profiles pointed out as potentially interesting (see section 5.4.1) are subjected to a deeper study based on data clustering techniques (see section 5.3). Here, the separation accuracies (see Table 5) are presented.

Table 5: Potentially interesting profiles and data clustering separation accuracies. (VC1: buildings; VC2: fire; VC3: rocks; VC4: smoke; VC5: snow; VC6: trees; VC7: water.)

Profile	SVM Acc. (%)	DC Accuracy (%) [VC1;VC2;VC3;VC4;VC5;VC6;VC7]
[CLD]	55.09	60.96 [73.74; 95.37; 39.00; 40.23; 64.95; 74.29; 39.13]
[SCD]	55.81	55.02 [44.40; 79.63; 48.00; 51.72; 71.13; 77.14; 13.04]
[CSD]	52.18	48.29 [43.434; 92.593; 61.00; 35.63; 42.27; 60.95; 2.17]
[SCD CLD]	56.4	55.87 [29.29; 97.22; 64.00; 60.92; 56.70; 59.05; 23.91]
[CSD CLD]	53.78	59.13 [65.657; 96.3; 40.00; 41.38; 65.98; 75.24; 29.39]
[SCD CLD HTD rSD]	64.24	57.61 [51.52; 96.30; 35.00; 54.02; 64.95; 74.29; 27.174]

Basing on the results of Table 5, it can be concluded that the use of the SVM (*Euclidean* distance) is acceptable for most of the cases, although in some the achieved accuracy is worse than for data clustering (i.e. [CLD] , because the proposed distance is neither *Euclidean* nor *Hamming*) . Of course, we cannot forget the fact that the compiled spatial regions do not represent all the space of images of the visual cues under study.

6. Conclusions & Future Work

In this work, a method for the estimation of the interdependency between classification accuracy (using a SVM with *Euclidean distance*) of spatial regions and computational cost of the required descriptor extraction has been proposed. Although the method has been applied to the concrete context of natural disasters, it can be used in other context as long as the appropriate content set is compiled. The reliability of the inferred conclusions is endorsed by the comparison with data clustering techniques.

We can conclude that, in the context under study, certain descriptors seem to be more recommendable for real time applications, since they improve the classification accuracy with very low computational cost increasing. A classification solution based on SVM is faster than data clustering techniques and so, it is recommendable when the visual descriptors work fine with the *Euclidean distance*. When more complex distances are proposed, SVM presents a decreasing in its separation capacity, but it keeps giving reasonable results, as this work corroborates.

One of the future action lines will consist on the study on temporal dependant descriptors, such as motion descriptors or shape evolution descriptors. It is also planned to apply some other spatial and spatiotemporal segmentation techniques in order to validate the classification results obtained on this work.

As already commented in section 3.2, the most common annotated ground-truth collections consist on manually annotated images but with no spatial information or segmentation. The work described in A.2 has not yet achieved conclusive results; it constitutes the current researching line. It faces the problematic of using a content set as the one described in section 3.2, a content set with annotations at image rather that spatial region level.

7. References

[1] Chang, S.-F.; Sikora, T.; Puri, A.; "Overwiew of the MPEG-7 standard", IEEE Transactions, Circuits Syst, Video Technol., 2001, 11, (6), pp. 688-695.

[2] Manjunath, B.S.; Salembier, P.; Sikora, T.; "Introduction to MPEG-7"; 1st edition. John Wiley & Sons, Ltd.; West Sussex, England.

[3] Mikolajczyk, K. ,Schmid, C., "A performance Evaluation of Local Descriptors", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 10, 2007, pp.1615 – 1630.

[4] Timo Ojala, Markus Aittola, Esa Matinmikko, "Empirical Evaluation of MPEG-7 XM Color Descriptors in Content-Based Retrieval of Semantic Image Categories", Proc. 16th International Conference on Pattern Recognition, 2002, Quebec, Canada, 2:1021-1024.

[5] Ramprasath Dorairaj, Kanesh R. Namuduri, "Compact Combination of MPEG-7 Color and Texture Descriptors for Image Retrieval", Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eigth Asilomar Conference on, Vol.1, pp. 387-391.

[6] Jorma Laaksonen, Markus Koskela, Erkii Oja, "PicSOM, Self-Organizing Image Retrieval with MPEG-7 Content Descriptors", IEEE Transactions on Neuronal Networks, Vol. 13, No. 4, pp. 841- 853, Jul. 2002.

[7] Evaggelos Spyrou, Hervé Le Borgne, Theofilos Mailis, Eddie Cooke, Yannis Avrithis, Noel O´Connor, "Fusing MPEG-7 Visual Descriptors for Image Classification", W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 847-852, 2005.

[8] Timo Ojala, Topi Mäenpää, Jaakko Viertola, Juha Kyllönen, Matti Pietikäinen; "Empirical Evaluation of MPEG-7 Texture Descriptors with A Large-Scale Experiment"; Proc. 2nd International Workshop on Texture Analysis and Synthesis, Copenhagen 2002, Denmark, 99 - 102.

[9] Bastian Leibe, Ales Leonardis, and Bernt Schiele; "Combined Object Categorization and Segmentation with an Implicit Shape Model"; ECCV'04 Workshop on Statistical Learning in Computer Vision, Prague, May 2004.

[10] Maha El Choubassi, Ara V. Nefian, Igor Kozintsev, Jean-Yves Bouguer, Yi Wu; "Web Image Clustering"; Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on Volume 4, 15-20 April 2007 Page(s):IV-1221 - IV-1224.

[11] R.Fergus, L.Fei-Fei, P. Perona, A. Zisserman; "Learning Object Categories from Google´s Image Search"; Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on Volume 2, 17-21 Oct. 2005 Page(s):1816 - 1823 Vol. 2.

[12] J.Winn, A. Criminisi, T. Minka; "Object Categorization by Learned Universal Visual Dictionary"; Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on Volume 2, 17-21 Oct. 2005 Page(s):1800 - 1807 Vol. 2.

[13] Sungho Kim, In So Kweon; "Simultaneous Classification and Visual Word Selection using Entropy-based Minimum Description Length"; Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06).

[14] Shi Rui, Wanjun Jin, Tat-Seng Chua; "A Novel Approach to Auto Image Annotation Based on Pair-wise Constrained Clustering and Semi-naïve Bayesian Model", Computational Intelligence in Image and Signal Processing, CIISP 2007, IEEE Symposium on. Pages: 301-306.

[15] Chiou-Ting Hsu, Chuech-Yu Li; "Relevance Feedback Using Generalized Bayesian Framework With Region-Based Optimization Learning"; Image Processing, IEEE Transactions on; Volume 14, Issue 10, Oct. 2005, Pages: 1617-1631.

[16] Russell, B.C., Torralba, A. Murphy, K.P. and Freeman, W.T. "LabelMe: a database and web-based tool for image annotation" MIT AI Lab Memo AIM-2005-025, September, 2005.

[17] TRECVID 2007 contest: http://www-nlpir.nist.gov/projects/tv2007/tv2007.html

[18] C.-Y. Lin, B. L. Tseng and J. R. Smith, "Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets,"NIST TREC-2003 Video Retrieval Evaluation Conference, Gaithersburg, MD, November 2003.

[19] Stéphane Ayache and Georges Quénot, "Evaluation of active learning strategies for video indexing", In Fifth International Workshop on Content-Based Multimedia Indexing (CBMI'07), Bordeaux, France, June 25-27, 2007.

[20] MPEG-7: Visual experimentation model (xm) version 10.0. ISO/IEC/JTC1/SC29/WG11, Doc. N4062 (2001).

[21] Dean S. Messing, Peter van Beek, James H. Errico, "The MPEG-7 Colour Structure Descriptor: Image Description using Colour and Local Spatial Information", Image Processing, 2001, International Conference on, Vol. 1, pp. 670-673.

[22] B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan and Akio Yamada, "Color and Texture Descriptors", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, June 2001

[23] Miroslaw Bober, "MPEG-7 Visual Shape Descriptors", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, June 2001

[24] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : a library for support vector machines", 2001. Software available at **http://www.csie.ntu.edu.tw/~cjlin/libsvm**

[25] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classi¯cation. 2 edn. Wiley Inter-science (2000).

[26] NTUA Colour Spatial Segmentator **http://www.image.ntua.gr**

[27] Matlab 7.0 R14, **http://www.mathworks.com/products/matlab/**

A. Appendix

A.1. Computational Cost and Classification Accuracy Relations

Profile	Comp. Cost (%)	Acc.(%)	Acc./C. Cost
[cSD]	4.38	21.66	4.9452
[rSD]	26.91	30.23	1.1234
[rSD cSD]	31.29	32.56	1.0406
[HTD]	36.92	49.56	1.3424
[HTD cSD]	41.3	48.84	1.1826
[HTD rSD]	63.82	51.02	0.79944
[HTD rSD cSD]	68.21	49.42	0.72453
[CLD]	0.9	55.09	61.211
[CLD cSD]	5.29	47.09	8.9017
[CLD rSD]	27.81	47.38	1.7037
[CLD rSD cSD]	32.2	48.26	1.4988
[CLD HTD]	37.82	60.32	1.5949
[CLD HTD cSD]	42.2	60.17	1.4258
[CLD HTD rSD]	64.73	59.59	0.92059
[CLD HTD rSD cSD]	69.11	58.58	0.84763
[CSD]	2.66	52.18	19.617
[CSD cSD]	7.04	51.89	7.3707
[CSD rSD]	29.57	49.27	1.6662
[CSD rSD cSD]	33.95	49.71	1.4642
[CSD HTD]	39.57	61.77	1.561
[CSD HTD cSD]	43.96	61.77	1.4051
[CSD HTD rSD]	66.48	61.48	0.92479
[CSD HTD rSD cSD]	70.87	60.17	0.84902
[CSD CLD]	3.56	53.78	15.107
[CSD CLD cSD]	7.95	54.8	6.8931
[CSD CLD rSD]	30.47	52.18	1.7125
[CSD CLD rSD cSD]	34.86	52.76	1.5135
[CSD CLD HTD]	40.48	63.37	1.5655
[CSD CLD HTD cSD]	44.86	63.23	1.4095
[CSD CLD HTD rSD]	67.39	61.05	0.90592
[CSD CLD HTD rSD cSD]	71.77	61.19	0.85258
[SCD]	2.73	55.81	20.443
[SCD cSD]	7.11	56.69	7.9733
[SCD rSD]	29.64	56.83	1.9173

[SCD rSD cSD]	34.02	56.69	1.6664
[SCD HTD]	39.64	63.66	1.606
[SCD HTD cSD]	44.03	63.95	1.4524
[SCD HTD rSD]	66.55	63.95	0.96093
[SCD HTD rSD cSD]	70.94	63.23	0.89132
[SCD CLD]	3.63	56.4	15.537
[SCD CLD cSD]	8.02	57.12	7.1222
[SCD CLD rSD]	30.54	57.12	1.8703
[SCD CLD rSD cSD]	34.92	56.83	1.6274
[SCD CLD HTD]	40.55	63.08	1.5556
[SCD CLD HTD cSD]	44.93	63.95	1.4233
[SCD CLD HTD rSD]	67.46	64.24	0.95227
[SCD CLD HTD rSD cSD]	71.84	63.95	0.89017
[SCD CSD]	5.39	56.25	10.436
[SCD CSD cSD]	9.77	57.41	5.8762
[SCD CSD rSD]	32.29	57.12	1.769
[SCD CSD rSD cSD]	36.68	56.69	1.5455
[SCD CSD HTD]	42.3	62.21	1.4707
[SCD CSD HTD cSD]	46.69	61.63	1.32
[SCD CSD HTD rSD]	69.21	60.47	0.87372
[SCD CSD HTD rSD cSD]	73.59	62.79	0.85324
[SCD CSD CLD]	6.29	56.69	9.0127
[SCD CSD CLD cSD]	10.68	56.98	5.3352
[SCD CSD CLD rSD]	33.2	57.41	1.7292
[SCD CSD CLD rSD cSD]	37.58	57.12	1.52
[SCD CSD CLD HTD]	43.21	62.21	1.4397
[SCD CSD CLD HTD cSD]	47.59	61.92	1.3011
[SCD CSD CLD HTD rSD]	70.11	60.32	0.86036
[SCD CSD CLD HTD rSD cSD]	74.5	62.79	0.84282
[DC1]	25.5	37.65	1.4765
[DC1 cSD]	29.89	38.81	1.2984
[DC1 rSD]	52.41	39.83	0.75997
[DC1 rSD cSD]	56.79	40.99	0.72178
[DC1 HTD]	62.42	54.22	0.86863
[DC1 HTD cSD]	66.8	54.36	0.81377
[DC1 HTD rSD]	89.32	55.23	0.61834
[DC1 HTD rSD cSD]	93.71	52.76	0.56301
[DC1 CLD]	26.41	44.77	1.6952

[DC1 CLD cSD]	30.79	44.77	1.454
[DC1 CLD rSD]	53.31	45.78	0.85875
[DC1 CLD rSD cSD]	57.7	46.8	0.81109
[DC1 CLD HTD]	63.32	57.12	0.90208
[DC1 CLD HTD cSD]	67.71	57.7	0.85216
[DC1 CLD HTD rSD]	90.23	58.87	0.65244
[DC1 CLD HTD rSD cSD]	94.61	57.7	0.60987
[DC1 CSD]	28.16	52.91	1.8789
[DC1 CSD cSD]	32.54	54.07	1.6616
[DC1 CSD rSD]	55.07	53.78	0.97658
[DC1 CSD rSD cSD]	59.45	55.38	0.93154
[DC1 CSD HTD]	65.08	61.19	0.94023
[DC1 CSD HTD cSD]	69.46	62.5	0.8998
[DC1 CSD HTD rSD]	91.98	61.05	0.66373
[DC1 CSD HTD rSD cSD]	96.37	61.34	0.63651
[DC1 CSD CLD]	29.06	54.65	1.8806
[DC1 CSD CLD cSD]	33.45	55.09	1.6469
[DC1 CSD CLD rSD]	55.97	54.94	0.9816
[DC1 CSD CLD rSD cSD]	60.36	55.38	0.9175
[DC1 CSD CLD HTD]	65.98	62.5	0.94726
[DC1 CSD CLD HTD cSD]	70.36	62.21	0.88417
[DC1 CSD CLD HTD rSD]	92.89	61.19	0.65874
[DC1 CSD CLD HTD rSD cSD]	97.27	61.05	0.62763
[DC1 SCD]	28.23	56.4	1.9979
[DC1 SCD cSD]	32.61	55.81	1.7114
[DC1 SCD rSD]	55.14	56.54	1.0254
[DC1 SCD rSD cSD]	59.52	56.98	0.95733
[DC1 SCD HTD]	65.14	62.06	0.95272
[DC1 SCD HTD cSD]	69.53	63.81	0.91773
[DC1 SCD HTD rSD]	92.05	63.81	0.69321
[DC1 SCD HTD rSD cSD]	96.44	64.1	0.66466
[DC1 SCD CLD]	29.13	56.54	1.941
[DC1 SCD CLD cSD]	33.52	56.4	1.6826
[DC1 SCD CLD rSD]	56.04	56.54	1.0089
[DC1 SCD CLD rSD cSD]	60.43	57.27	0.94771
[DC1 SCD CLD HTD]	66.05	61.77	0.9352
[DC1 SCD CLD HTD cSD]	70.43	63.66	0.90388
[DC1 SCD CLD HTD rSD]	92.96	63.37	0.68169

[DC1 SCD CLD HTD rSD cSD]	97.34	63.95	0.65698
[DC1 SCD CSD]	30.89	57.12	1.8491
[DC1 SCD CSD cSD]	35.27	57.41	1.6277
[DC1 SCD CSD rSD]	57.8	56.83	0.98322
[DC1 SCD CSD rSD cSD]	62.18	57.7	0.92795
[DC1 SCD CSD HTD]	67.8	61.77	0.91106
[DC1 SCD CSD HTD cSD]	72.19	62.65	0.86785
[DC1 SCD CSD HTD rSD]	94.71	62.06	0.65526
[DC1 SCD CSD HTD rSD cSD]	99.1	62.35	0.62916
[DC1 SCD CSD CLD]	31.79	57.7	1.815
[DC1 SCD CSD CLD cSD]	36.18	57.85	1.5989
[DC1 SCD CSD CLD rSD]	58.7	56.98	0.9707
[DC1 SCD CSD CLD rSD cSD]	63.08	58.14	0.92169
[DC1 SCD CSD CLD HTD]	68.71	61.48	0.89478
[DC1 SCD CSD CLD HTD cSD]	73.09	62.35	0.85306
[DC1 SCD CSD CLD HTD rSD]	95.62	61.63	0.64453
[DC1 SCD CSD CLD HTD rSD cSD]	100	62.06	0.6206
[DC2]	21.95	36.92	1.682
[DC2 cSD]	26.34	37.94	1.4404
[DC2 rSD]	48.86	38.37	0.7853
[DC2 rSD cSD]	53.24	40.26	0.7562
[DC2 HTD]	58.87	54.07	0.91846
[DC2 HTD cSD]	63.25	53.63	0.84791
[DC2 HTD rSD]	85.77	54.65	0.63717
[DC2 HTD rSD cSD]	90.16	53.05	0.5884
[DC2 CLD]	22.86	43.46	1.9011
[DC2 CLD cSD]	27.24	44.77	1.6435
[DC2 CLD rSD]	49.76	44.77	0.89972
[DC2 CLD rSD cSD]	54.15	45.49	0.84007
[DC2 CLD HTD]	59.77	57.27	0.95817
[DC2 CLD HTD cSD]	64.16	57.41	0.89479
[DC2 CLD HTD rSD]	86.68	58.72	0.67743
[DC2 CLD HTD rSD cSD]	91.06	57.27	0.62893
[DC2 CSD]	24.61	52.18	2.1203
[DC2 CSD cSD]	28.99	53.05	1.8299
[DC2 CSD rSD]	51.52	52.91	1.027
[DC2 CSD rSD cSD]	55.9	54.65	0.97764
[DC2 CSD HTD]	61.53	61.05	0.9922

[DC2 CSD HTD cSD]	65.91	61.92	0.93946
[DC2 CSD HTD rSD]	88.43	61.48	0.69524
[DC2 CSD HTD rSD cSD]	92.82	60.9	0.65611
[DC2 CSD CLD]	25.52	53.49	2.096
[DC2 CSD CLD cSD]	29.9	53.92	1.8033
[DC2 CSD CLD rSD]	52.42	53.49	1.0204
[DC2 CSD CLD rSD cSD]	56.81	54.36	0.95687
[DC2 CSD CLD HTD]	62.43	62.35	0.99872
[DC2 CSD CLD HTD cSD]	66.82	61.92	0.92667
[DC2 CSD CLD HTD rSD]	89.34	61.19	0.68491
[DC2 CSD CLD HTD rSD cSD]	93.72	60.47	0.64522
[DC2 SCD]	24.68	55.81	2.2613
[DC2 SCD cSD]	29.06	55.96	1.9257
[DC2 SCD rSD]	51.59	56.54	1.0959
[DC2 SCD rSD cSD]	55.97	56.98	1.018
[DC2 SCD HTD]	61.6	62.06	1.0075
[DC2 SCD HTD cSD]	65.98	63.52	0.96272
[DC2 SCD HTD rSD]	88.5	63.95	0.7226
[DC2 SCD HTD rSD cSD]	92.89	63.95	0.68845
[DC2 SCD CLD]	25.58	55.81	2.1818
[DC2 SCD CLD cSD]	29.97	56.1	1.8719
[DC2 SCD CLD rSD]	52.49	56.69	1.08
[DC2 SCD CLD rSD cSD]	56.88	57.27	1.0069
[DC2 SCD CLD HTD]	62.5	61.77	0.98832
[DC2 SCD CLD HTD cSD]	66.88	63.37	0.94752
[DC2 SCD CLD HTD rSD]	89.41	63.66	0.712
[DC2 SCD CLD HTD rSD cSD]	93.79	64.1	0.68344
[DC2 SCD CSD]	27.34	57.41	2.0999
[DC2 SCD CSD cSD]	31.72	56.98	1.7963
[DC2 SCD CSD rSD]	54.25	56.98	1.0503
[DC2 SCD CSD rSD cSD]	58.63	57.85	0.9867
[DC2 SCD CSD HTD]	64.25	61.92	0.96374
[DC2 SCD CSD HTD cSD]	68.64	62.65	0.91273
[DC2 SCD CSD HTD rSD]	91.16	62.06	0.68078
[DC2 SCD CSD HTD rSD cSD]	95.55	62.21	0.65107
[DC2 SCD CSD CLD]	28.24	57.56	2.0382
[DC2 SCD CSD CLD cSD]	32.63	57.41	1.7594
[DC2 SCD CSD CLD rSD]	55.15	56.83	1.0305

[DC2 SCD CSD CLD rSD cSD]	59.54	58.14	0.97649
[DC2 SCD CSD CLD HTD]	65.16	61.63	0.94583
[DC2 SCD CSD CLD HTD cSD]	69.54	62.21	0.89459
[DC2 SCD CSD CLD HTD rSD]	92.07	61.63	0.66938
[DC2 SCD CSD CLD HTD rSD cSD]	96.45	61.77	0.64044

A.2. Current status of Image Separation work

A.2.1 Proposed Method

Here we propose a method that, beginning from a basic colour spatial segmentation [26] with a fixed number of output regions, eight, performs spatial regions clustering. This region merging uses as elements´ patterns the visual descriptors enumerated in section 4.1 and as distances, the dissimilarity measures described in section 4.2. Some implementation difficulties are found when trying to directly cluster this elements (see A.2.2), for solving this, a semantic guided clustering is proposed (see A.2.3).

The similar regions (in terms of visual descriptions) join in clusters which, depending on the number of components of the clusters and on the entropy rate of the semantic annotations of the images they belong to, present different reliability in its separation results (see A.2.4).

Based on the conclusions presented in section 5.4 and on the intrinsic limitations of the low-level clustering methods used on this study (see A.2.3), different combinations of visual descriptors (see section 4) are used as patterns.

In the Figure 4 we can see and example of the proposed semantic guided low level clustering and degree of truth estimation of the clusters.

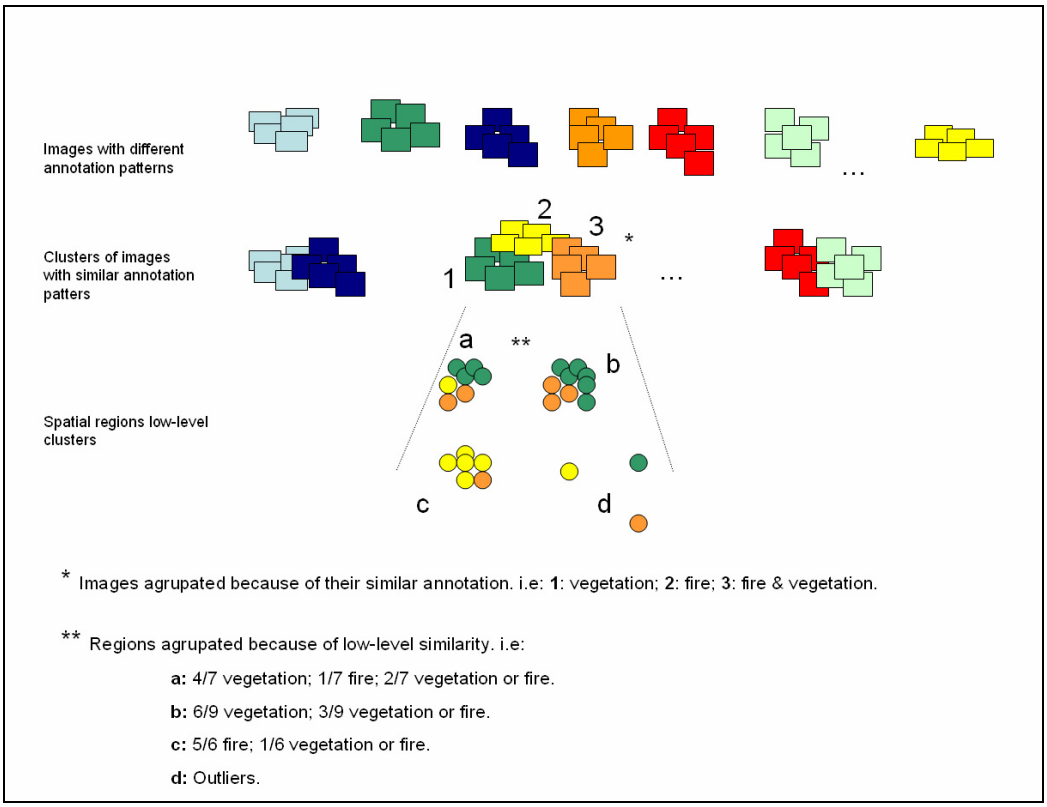

Fig. 4. Semantic guided low-level clustering example

A.2.2 Visual Descriptions and Data Clustering restrictions

Data clustering algorithms, in general, have a very high computational cost in terms of processing time and memory [25]. Concretely, when using hierarchical clustering, the distances between all the pair of elements are calculated. Using this algorithm, for a space of

n elements, $\frac{n(n-1)}{2}$ distances need to be processed. When using the k-means algorithm, although the processing is lighter, the number of elements in our case is still too high.

The use of the content set described in section 3.2 results in a space of around 170000 elements to be clustered. Such a huge number of elements make the direct application of a hierarchical or even a k-means clustering unfeasible.

A.2.3 Data Clustering Guided by Semantic information

So, in order to face the unfeasibility of performing a data clustering over the set of elements under study, the following approach is proposed. The low-level features clustering is performed over the groups of elements created following a semantic criterion, this is, based on the manual annotation available (see section 3.2). The process consists on a loop with the following two execution steps:

1. Semantic clustering of images: Its implementation consists of a k-means clustering using *cityblock* distance and a random sample of the input elements (the number of seeds changes on each iteration of the loop and depends on the low-level clustering method being used). The patterns for this merging are created as 39 length vectors in which each position represents each of the visual cues annotated (see section 3.2). Their possible values are: 0, *negative*, 1; *skipped* or *not annotated*; 2, *positive*. So, this way we are joining images with similar manual annotations.
2. Low-level clustering of elements: It is applied to each of the groups resulting of the first step of the loop. The groups being clustered are composed of spatial regions (their low-level descriptions) which belong to images with similar annotations (to the same semantic cluster). This is a way of guidance of the low-level clustering, since the semantic groups tend to concentrate regions representing the same visual cues.

With each iteration of the loop the number of elements decreases, since some of them get represented by the centroids of the clusters to which they are added. This allows the reduction of the number of the semantic groups in the following iteration, because their size is not as big as originally was. The loop stops when all the elements (more precisely, the representatives of all the original elements) are subjected, together, to a low-level clustering.

The sequence of numbers of semantic clusters used on each iteration depends on the low-level clustering method being applied, since each method (and of course, each implementation) is able to work with different number of elements:

- K-means: This approximation consists on the appliance of the k-means algorithm to the groups resulting of the semantic clustering. As already commented, the final aim is to cluster a large amount of elements, for what this are divided in groups following a semantic criterion. Considering the concrete implementation [27] we are working with, the maximum number of processable elements is around 15000.
- Hierarchical Clustering: Hierarchical technique is able to work with less number of elements than K-means does [27], consequently, it will need more repetitions of the described loop in order to perform a complete low-level clustering.

A.2.4 Degree of Truth of the Clusters.

Once the spatial regions are clustered basing on low-level features, each of the clusters is characterized with the semantic information of the images the regions (is composed by) belong to. This characterization consists on the sum of the semantic annotation patterns (see A.2.3) of the images of origin of the cluster components (spatial regions). The higher the value of a certain position of the sum vector is, the more probable the occurrence of the visual

cue (see section 3.2.1) associated to that position is. Now on, this vector is named: $v = \sum annotations$.

The reliability of the probability estimations of a cluster depends of the number of components of the cluster and on the value of the entropy rate calculated as follows:

$H(X) = -\sum_{i=1}^{Ncues} p_i \log_2 p_i$, where: *Ncues* is the number of visual cues manually annotated in our content set (see section 3.2.1) and $p_i = \frac{v(i)}{\sum_{\nabla j} v(j)}$.

This way, the clusters with low number of components or low value of entropy are less reliable than the ones that, with an elevated number of components, keep a high entropy rate.