Linking visual cues and semantic terms under specific digital video domains

Juan M. Sánchez, Xavier Binefa, Jordi Vitrià and Petia Radeva

Computer Vision Center and Departament d'Informàtica
Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
e-mail: {juanma, xavierb, jordi, petia}@cvc.uab.es

Abstract

Semantic retrieval from video databases is becoming a very important research topic in the area of multimedia. This kind of tasks require the development of video data representation models which include the relationships between low level visual cues and the semantic concepts inferred from them. This paper presents a work based on semiotic studies that includes the extraction of simple visual features from commercials and a statistical analysis of them and their relationships with high level semantic terms. Well-known algorithms have been implemented and enhanced for feature extraction, as well as a novel probabilistic approach to color naming. The statistical analysis consists of finding correlations between variables, as well as the dimensions in feature space that best explain the variance of the data set. Some interesting conclusions are reached at the end of the work about how commercials are grouped in feature space with respect to different levels of semantics.

1   Introduction

The large amount of information that is being compiled in the everyday faster growing number of digital video libraries that can be found all over the world has directed the efforts of many researchers to the development of efficient ways of indexing their contents in order to accomplish subsequent retrieval tasks [2, 5, 7]. In order to give support to what is known as high-level or semantic retrieval, adequate models for describing video content must be developed. The heterogeneity of the content of video in the general sense leads to think that there is no one general model, so that different models, which could cooperate in a general purpose framework, should be developed for different kinds of contents [11].

Picard suggested in [12] what possibly the key challenges within this research topic are: finding good models for describing visual content in the sense of both compression and content access and finding good measures of visual similarity. A model for describing video content, while giving support for both perceptual and semantic retrieval, consists of: (1) a set of meaningful low level visual cues, (2) links between them and

high level semantic concepts and (3) a similarity criterion (or several) that fits the one used by humans, since the final users will ultimately evaluate the performance of the model in the task of content retrieval from digital video libraries.

Semiotics is a science of great interest when it comes to developing models for specific domains. The semiotic perspective of visual video content can help us to determine several parts of the model, starting from the selection of meaningful visual cues, up to their relationships with semantic terms and the choice of similarity criteria. The work presented in this paper is directed to the domain of TV commercials. From the semiotic perspective, the members of this domain have certain regularity patterns that lead to a semantic classification. There are plenty of evidences that suggest the use of color, orientation and motion as visual cues, together with edit rhythm information, as well as evidences of how all this data is related to the sensations that arise in the viewer from their use. This paper goes in for the extraction of all this low level information from commercials and its analysis in order to find relationships with high level terms and if they adequately fit the relationships established by human perception.

The rest of the paper is organized as follows. Section 2 gives a brief introduction to semiotics and the codes used in commercials. The extraction of simple visual cues from the videos is treated in section 3 and a statistical analysis of the data obtained from a set of commercials is detailed in section 4. Finally, the paper is concluded in section 5.

2   Semiotics in the TV advertising domain

Semiotics is the science that studies the relationships between signs and their meanings, that is to say *codes*. Within the semiotics perspective, these relationships are purely arbitrary and they are imposed by social conventions in a specific cultural context. Visual semiotics works on the codes that make people understand images, and takes out from questions closely related to computer vision such as how 3D worlds are recovered from flat surfaces, or if images are read in the same way a text is [17].

The point of view of semiotics is of great interest to our work in order to allow us to know how visual information is related to the semantics of images and video. Colombo et al. introduced in [4] and mainly in [3] the work of semiotics in the domain of commercials, or what they called the *language* of commercials. In their work, they gave a set of *a priori* relationships between visual cues such as color, camera work and edit effects, and their associated semantics. The semantic level can be seen in two different ways: the feelings arosen in the viewer by the images (happiness, relax, suspense, stirring, ...), and what has been called *consumption values* (practical, utopic, critical and playful) which are directly related to the four basic narrative elements found in the structure of commercials [3]. Some details about the *language* of commercials are briefly summarized next.

Concerning to color, warm ones grab the attention of the observer and communicate dynamism. On the other hand, cold colors suggest gentleness, calm, relax and faithfulness. Each different color can be associated to a set of feelings, which are summarized in table 2. In a different way, the use of saturated colors is a sign of less realistic situations than using unsaturated ones, giving a sense of fancy and joyful worlds and

Color	Feelings evoked in the viewer
Red	Happiness, dynamism, aggresiveness, violence, power
Orange	Glory, solemnity, vanity, progress
Golden yellow	Richness, prosperity, happiness
Dark yellow	Deception, caution
Green	Calm, relax, hope
Blue	Gentleness, fairness, faithfulness, virtue
Purple	Melancholy, fear
Brown	Relax (mostly used as background color)

Table 1: Summary of the feelings related to colors.

communicating happiness. The presence of light colors also induces the viewer to feel calmed and relaxed.

Camera work and framing is closely related to the amount of action and dynamism in the video. Fast motion of the camera and the actants, i.e. any person or object that performs an action, communicates dynamism and stirring, while still takes mean quietness, calm and relax. The slope of the scene is also considered. Slanted slopes communicate action, but also happiness and they want to evoke a degree of unreality in the scene, while horizontal and vertical slopes communicate calm. It is also interesting to note that slopes in the scene can be used as an heuristic to determine if it has been shot in a man-made scenery, where a dominant slope is found, mainly horizontal or vertical, or in a natural one, where there is not any dominant slope.

Edit rhythm is also directly related to the amount of action and dynamism that the advertiser wants to communicate to the viewers. A video edited using few long shots and gradual transitions inspires quietness, while videos with many short shots using cuts emphasize dynamism and happiness.

3   Extracting visual cues

In this section the algorithms used for extracting meaningful low level visual cues are reviewed and analyzed. All measures are considered from a probabilistic point of view, i.e. ranging from 0 to 1 and having certain dependencies between them. Semiotics, as stated in the previous section, suggests the use of the following video features as a first step: the presence of each different color, dominant slopes, the amount of motion and the dynamism imposed during editing.

3.1   Color

The measures of color we are interested in are the probabilities that given a pixel of the video, it belongs to each color defined by semiotics to be relevant. Therefore, given the coordinates of a pixel in a particular color space, it must be related to predefined color classes or labels. This leads us to the formulation of the color naming problem.

3

3.1.1 Probabilistic color naming

Lammens defined color naming in [9] as a mapping N from visual stimuli to pairs of color terms, $c_i \in C$, and "membership" or "confidence" measures. At first, visual stimuli are represented by spectral distributions, but computationally speaking, the co-ordinates of the pixels in a color space are used. The color terms to be used should be those suggested by semiotics, but the algorithm implemented in our work involves a general purpose probabilistic classification framework, so that the color classes are defined during the model training step. The probabilistic point of view gives sense to the concept of confidence in a natural way.

The problem of color naming can be formulated as a classification problem, where the input variables are coordinates in a color space and the output is the color term or class. As a classification problem, it could be faced using many different techniques. The probabilistic classification approach seems to be the most suitable in this case, since a value of membership is required. Probabilistic classification consists of assigning a model to each color class, i.e. a probability distribution, so that the mapping would be:

$$N(\vec{x}) = <c_i, P(c_i|\vec{x})> \tag{1}$$

where $i = \arg\max P(c_i|\vec{x})$, i.e. the color is assigned to the label with the highest probability, resulting in a maximum a posteriori (MAP) classifier. These probabilities are computed using Bayes' rule as:

$$P(c_i|\vec{x}) = \frac{P(\vec{x}|c_i)P(c_i)}{\sum_i P(\vec{x}|c_i)P(c_i)} \tag{2}$$

$P(\vec{x}|c_i)$ is the likelihood of the color $\vec{x}$ given that the color model is c_i and $P(c_i)$ is the prior probability of this model. When all classes have the same prior probabilities, it is equivalent to finding the maximum likelihood (ML) class.

In the regular color naming problem, the input vector $\vec{x}$ should be assigned to the color label with the highest membership probability, but in our application it is pretty interesting to have a measure of the membership probabilities of every color class, given color $\vec{x}$. Our final feature vector corresponding to color will be the expected posterior probabilities of each color label, given a pixel from the video, that is to say the relevance of each color term in the video. Therefore, our mapping is a generalization of the one in eq. (1), omitting the final classification step:

$$N(\vec{x}) = \{<c_i, P(c_i|\vec{x})>, \forall c_i \in C\} \tag{3}$$

And the final feature vector for color has the following components:

$$f_i = P(c_i|\mathcal{X}) = \frac{1}{M} \sum_{m=1}^{M} P(c_i|\vec{x}^{(m)}) \tag{4}$$

where $\mathcal{X}$ is the set of the M pixels in the whole video.

Two main considerations must be done when using this algorithm: (1) the probability distribution we will be using to model each color class, and (2) the color space we will be working in.

When considering simple color names, like black, white, red, green and so on, a simple Gaussian model performs as a good estimation of the distribution of each color. But when introducing complex color terms like skin-color, which has been proved to be very useful and widely used for face detection and tracking, more complex mixtures of distributions, usually Gaussians, are used due to non-linearities in the distribution that are not captured by simple Gaussian models. Thus, the likelihood of having a color $\vec{x}$ given a particular color model c_i would be:

$$P(\vec{x}|c_i) = \sum_j \pi_{ij} \frac{1}{\sqrt{(2\pi)^{d/2}|\Sigma^{(i^j)}|}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}^{(i^j)})^T \left[\Sigma^{(i^j)}\right]^{-1}(\vec{x} - \vec{\mu}^{(i^j)})\right\}$$

(5)

where the index i^j stands for Gaussian j of the mixture corresponding to color class i. Particular cases of this approach can be found in the literature, mostly devoted to skin-color segmentation ([19]).

With respect to the choice of color space, at a first glance, the algorithm is not dependent of the color space used. A probability distribution that fits each color class can be found in any color space, so that RGB can be a good choice in order to avoid color space conversions. Even though, the use of a normalized chromatic color space like rg is not adequate in the general case due to the impossibility to distinguish between light and dark colors, black and white and so on. Nevertheless, the choice of color space or a comparative analysis of the performance of the algorithm in different spaces is out of the scope of our work and will not concern us.

3.1.2 Estimating color model parameters

The parameters of the probability density function that models a particular color class are estimated from training data. The model learning process can be supervised or unsupervised.

Supervised For this training procedure, it is necessary to have a set of data samples which classification is known. These sets were obtained by offering a sequence of images to the user, who was asked to select pixels that he/she considered to belong to each proposed color class. Then, assuming a 3D Gaussian distribution $G(\vec{\mu}, \Sigma)$ in RGB color space, its parameters for class c_i can be estimated from a set of N_i color samples $\vec{x}^{(j)}$ as:

$$\vec{\mu}^{(i)} = \frac{1}{N_i} \sum_{j=1}^{N_i} \vec{x}^{(j)} \tag{6}$$

$$\Sigma^{(i)} = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (\vec{x}^{(j)} - \vec{\mu}^{(i)})(\vec{x}^{(j)} - \vec{\mu}^{(i)})^T \tag{7}$$

In the case of having a complex distribution of the training samples that is modeled by a mixture of Gaussians, the well-known EM algorithm is used to obtain the parameters π_{ij}, $\vec{\mu}^{(i^j)}$ and $\Sigma^{(i^j)}$ of the distribution for class c_i [6].
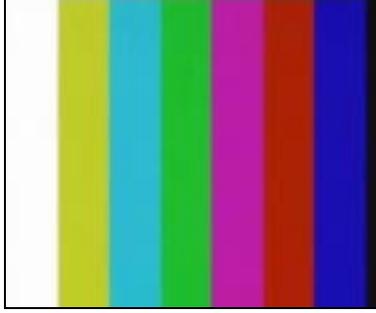
5

Figure 1: Color calibration pattern generated by a video device.

The number of samples selected by the user in the different training data sets is not imposed, so that the size of these sets may vary from one to the others. This fact was taken into account in the sense of prior probabilities of each class, which were computed from the whole training data as:

$$P(c_i) = \frac{1}{N} \sum_n P(c_i | \vec{x}^{(n)}) \tag{8}$$

As this is a supervised procedure, the probability of label c_i given $\vec{x}^{(n)}$ is known to be 1 if it belongs to the ith training set and 0 otherwise.

Unsupervised In this case, the parameters of the probability distributions which model the different color classes are estimated from a data set whose classification is unknown or hidden. Therefore, this is a suitable problem to be solved using the EM algorithm, particularly in the case of simple Gaussian models where a mixture of Gaussians is fitted on the full training data set and each distribution of the mixture corresponds to a different color class. One of the main problems of using EM is its dependence on initialization. If we initialized the parameters randomly and let the algorithm iterate, the resulting classes would very probably not agree with the color classes we are considering. Fortunately, some more information can be considered in order to guide the algorithm to the desired solution. To be exact, information about the location of the centroids of the classes in color space is provided by some kind of calibration procedure. There are several video devices able to generate color calibration patterns like the one shown in fig. 1 [1]. Thinking about simple colors, like the ones suggested by semiotics to be important, it is a good choice to consider the location of pure colors as given by the calibration pattern as the centroids of simple Gaussian models. Therefore, the parameters $\vec{\mu}^{(i)}$ are known and fixed, so that only $P(c_i)$ and $\Sigma^{(i)}$ are to be estimated by EM.

Clearly, the models obtained by means of an unsupervised training can not be as accurate as the ones obtained by a supervised technique. Accuracy can be improved by relying on the user to correct possible mistakes of the algorithm, thus guiding the

---

[1]Full color images from this and following figures will be available via WWW.

algorithm to a better model estimation. User intervention is also interesting in order to reduce the number of samples in the training data set by selecting a representative set of image pixels. In this case, we would be working on a mixed approach between supervised and unsupervised learning.

Some results comparing both training approaches are shown in fig. 2. The pixels in the resulting images are colored using their MAP classification. The maximum posterior probabilities are also shown. Note that color labeling using supervised training is much more accurate and realistic than using the unsupervised one. Furthermore, the figure shows how the models obtained via unsupervised learning are free to capture any training points, provided that global likelihood of the data is maximized (see that gray is considered yellow with unsupervised learning and white with supervised, which is much more realistic). Probability maps are also useful in the sense of "suggesting" wrong classifications. Fig. 3 shows the labeling using supervised training of a color gradient from blue to green, where the probability of a true classification is relatively high in plain regions, but it is much slower during the transition.

As a summary, we should say that our feature vector corresponding to color is obtained by probabilistic means. In our particular case, the color classes considered are $C = \{Black, White, Red, Green, Blue, Cyan, Yellow, Magenta\}$, i.e. the corners of the RGB color space cube. Each color class is modeled by a probability distribution, whose parameters can be estimated in supervised or unsupervised way. Then, for any given pixel, its color $\vec{x}$ can be assigned a set of class memberships computed from the posterior probabilities of each class. In this way, the final feature vector obtained represents the expected posterior probabilities of each color class or label, given a pixel in the video.

3.1.3 Intensity and Saturation

The measures for intensity and saturation are computed from the corresponding coordinates in the HSI color space. The conversion from RGB values is straightforward:

$$I \quad = \quad \max\{R, G, B\} \tag{9}$$

$$S \quad = \quad \frac{\max\{R, G, B\} - \min\{R, G, B\}}{\max\{R, G, B\}} \tag{10}$$

The I and S values are averaged over the total number of pixels in the video and normalized in the range $[0, 1]$. In this way, they can be seen as the probabilities of the video having bright and/or saturated colors.

3.2 Orientation

In this case, the selected measures are the probabilities of horizontal, slanted and vertical orientations being the most outstanding in the video. For their computation, the Hough transform has been used, enhanced with information about the gradient of the lines. Recall that this transform is intended for detecting lines (or other parametric

Figure 2: Comparison of probabilistic color naming using supervised and unsupervised training. For each image, from left to right and top to bottom: the original image, color labeling with supervised learning, with unsupervised learning, and the likelihood maps of the classifications being true using supervised and unsupervised learning.

Figure 3: Meaning of the label probability maps seen in the case of a color gradient. The original image (left), its labeling using supervised training (middle) and the probabilities of this labeling (right) are shown.

Figure 4: Example of the Hough transform enhanced with gradient information. From left to right: the original synthetic image, its thresholded gradient for edge detection, its regular Hough transform, and its enhancement including information about the gradient. Accumulation space is formed by θ (ranging $[0, \pi]$) and ρ as horizontal and vertical axis (see text for the interpretation of different angle values). The two vertical lines are clearly detected by the Hough transform, but they can be vaguely perceived with respect to the slanted one in the enhanced version.

models) in edge images. In the case of straight lines, in order to avoid infinite slopes, the parameterization is usually done with respect to an angle θ and its distance ρ to the origin:

$$\rho = x \cos \theta + y \sin \theta \tag{11}$$

Then, for every edge point (x, y) of the image, a vote is given to every pair (θ, ρ) that fulfils equation 11. More information about the Hough transform and its generalized version can be found in the literature [1].

For our application, the algorithm has been enhanced in order to take into account information about the gradient of the image in edge points. In this way, the detection of the most outstanding lines is considering two aspects: their size and their contrast. Then, instead of suming one vote to every point that fulfils equation 11 in the accumulation space, the gradient modulus in (x, y) is added. An example on a synthetic image is shown in fig. 4, where the relevance of the two vertical lines is clearly reduced with respect to the slanted one, whose contrast is much higher, when considering their gradients. Its usefulness when working on real world imagery is also exemplified in fig. 5.

In order to obtain the final probabilities of horizontality, verticality and inclination, different ranges of angles are considered. All detected lines that fall in these intervals are sumed to compute the measures. As the angle considered is the one corresponding to the orthogonal line, the ranges are $[-20, 20]$ for vertical lines, $[70, 110]$ for horizontal ones, and $[20, 70] \cup [110, 160]$ for slanted. Some kind of fuzzy functions could have been used instead of straight intervals in order to get a better estimation, but the error caused by the simplest approach is allowable because the measures are averaged over all the frames of the video.
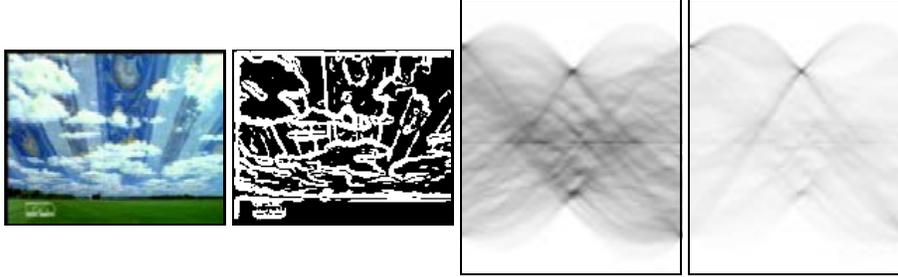
9

Figure 5: The same example as in fig. 4, but in the case of a real world image. Note that most of the irrelevant orientations are not significatively represented in the enhanced Hough transform, while horizontality ($\theta \approx \pi/2$ in accumulation space) is very strong due to the horizon line between land and sky.

3.3   Motion

The simplest algorithm for detecting and estimating motion is image differentiation [15]. The difference image I_d is computed straightforward by a simple pixel-wise subtraction:

$$I_d(x, y, t_i, t_{i+1}) = I(x, y, t_{i+1}) - I(x, y, t_i) \tag{12}$$

Concerning our work, the most important fact to be considered is that difference images can be used as a simple 2-point finite difference approximation of the temporal derivative of the image function $\partial I(x, y)/\partial t$ at the midpoint of the time interval $[t_i, t_{i+1}]$.

The amount of motion between two consecutive frames can be obtained from the number of changed pixels by thresholding the difference image. Again, this measure is normalized with respect to the total number of pixels in a frame in order to get a value in the range $[0, 1]$ that can be interpreted as the probability of having fast motion between frames, and it is then averaged over all the frames in the video.

3.4   Edit rhythm

Edit rhythm can be expressed both as a shot change rate or an average shot length. Many different algorithms are commonly used for detecting scene breaks, from the simplest color histogram differentiation algorithms [7], to Zabih's edge-based one [20] or our shot segmentation algorithm which is based on the variation of edges and the color around them in consecutive frames [14].

In this sense, there is not an immediate way of computing a measure following the probabilistic framework we are working on. The probability of finding a shot change in the video can be obtained as the ratio of scene breaks over the total number of frames. Although this is a good measure in the sense of probability, even in a high edit rhythm video, an average of 1 shot per second can be found, which means a shot change probability of 0.04. Therefore, the actual scale of this measure is too low with

respect to the others and this fact should be taken into account during the rest of the study.

Other possible features to be taken into account might be the ratio of sharp vs. smooth transitions. A scene break detector able to mark out different kinds of transitions, like the one in [14], is needed to automatically obtain this measure.

3.5 Other feasible features

There are many other features that could be considered in order to get a semantic interpretation as suggested by semiotic studies. They have not been focused by our work, but some links to the literature are given in this section.

For example, we can think about the presence of people during the video, even distinguishing between close-ups, full body shots and crowds. Many efforts have been devoted to research on face detection algorithms, including neural network [13], filtering [16] and eigenspace [8] approaches. All these algorithms can be improved by reducing the search space using skin-color segmentation (see section 3.1). The presence of crowds can also be detected by combining skin-color detectors and texture analysis. Other work on people detection in static and video images was presented by Papageorgiou et al. in [10].

Motion information could be enhanced in order to obtain different statistics for camera and object motions. Xiong and Lee presented in [18] an optical flow based approach to detecting and classifying different camera operations, like pan, tilt or zoom. Having estimated camera motion, object motions can also be obtained.

Finally, regarding to edit rhythm, other measures to be considered are the ratio of long and short shots. The problem here is when to consider a shot to be long or short.

4 Analysis of observed data

For our study, we have selected a set of 17 features for each commercial: 8 related to color considering the probabilities of finding black, white, red, green, blue, cyan, yellow and magenta pixels in the video, global intensity and saturation, 3 related to the probabilities of horizontal, slanted and vertical orientations, a measure of the global amount of motion, the rate of transitions per frame, and the ratios of cuts and dissolves with respect to the total number of transitions. From now on, these variables will be identified in the text following table 2. These measures were obtained using the algorithms from section 3. 70 commercials have been chosen for our preliminary statistical analysis of these data, which consists of the following inquiries:

- Analyze possible correlations between variables.

- Find the true dimensionality of the data, as well as the dimensions that generate the subspace of commercials.

- See if commercials are grouped in feature space (or in factor space) in the same way they are semantically grouped by humans.

Var. name	Id.	Var. name	Id.
Black	B	White	W
Red	R	Green	G
Blue	BL	Cyan	C
Yellow	Y	Magenta	M
Intensity	I	Saturation	S
Horizontal	H	Slanted	SL
Vertical	V	Motion	MO
Edit rhythm	ER	Cut ratio	CR
Dissolve ratio	DR		

Table 2: Variable identifiers.

The commercials selected for the study were digitally acquired from commercial blocks broadcast by different Spanish TV channels, without the purpose of finding a *corpus* of the most significant videos for our study (this may remain for a further work).

4.1 Correlation analysis

For the analysis of the correlations between variables, the correlation matrix R is used, computed from the data matrix X, where each row is a case, by means of Pearson's correlation coefficient as:

$$R_{ij} = \frac{\sum_k (X_{ki} - \mu_i)(X_{kj} - \mu_j)}{(N-1)\sigma_i \sigma_j} \tag{13}$$

where μ_i and σ_i are the mean and standard deviation of variable i. Apart from the obvious correlation between CR and DR, there are two more pairs of variables which are good *a priori* candidates to be highly correlated: B-I and MO-ER. The presence of black pixels is closely related to low intensity in the images because their luminance is almost null. On the other hand, a large amount of motion in the scenes means that there is action and dynamism, which is also inferred from a high rate of shot changes per frame. These relationships can be visually observed using scatterplots, as shown in figure 6.

The observation of the plots shows that there is a high correlation between B and I, but not between MO and ER. This means that the sensation of dynamism is not always inferred from both features, but it can be evoked by only one of them. As an example, action is communicated in Latin American serials mainly by object and people motion, but the rate of scene breaks is usually kept low. On the other hand, North American and European tendencies make use of the combination of both techniques, stressing shot changes. Some other correlations can be found analyzing the correlation matrix, which is shown in table 3. Note also some high correlations, which could be rather predicted, between B and W or W and I, and some others between S-R, S-Y and S-G, which lead us to think that saturated images are mostly composed of red, yellow and green.
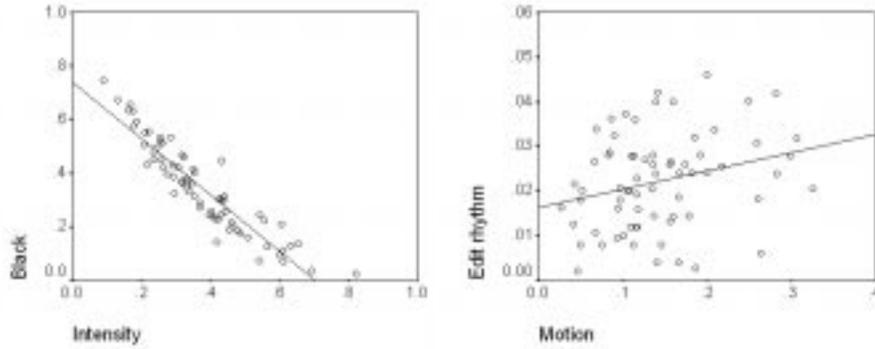
12

Figure 6: Scatterplots of B-I (left) and MO-ER (right), showing the regression line interpolating the data.

R	W	R	G	BL	C	Y	M	I	S	H	SL	V	MO	ER	CR	DR
B	-.557	.009	-.192	-.128	-.124	-.437	-.210	-.927	-.348	.041	.349	-.507	-.146	.056	.042	-.042
W		-.338	-.280	-.143	-.124	-.193	-.130	.590	-.485	.133	-.404	.339	-.088	-.140	-.129	.129
R			-.063	-.103	-.219	.294	.662	-.189	.606	-.304	.344	-.027	.204	-.047	-.232	.232
G				.206	.223	-.072	.123	.123	.403	-.042	.134	-.116	.164	.074	.066	-.066
BL					.509	-.182	.200	.015	.303	-.010	.154	-.185	.081	.077	.135	-.135
C						-.336	-.040	.159	.134	.197	-.115	-.121	.029	.121	.146	-.146
Y							.132	.392	.568	-.213	-.102	.423	.121	.004	.040	.040
M								.036	.485	-.201	.267	-.070	.279	-.060	-.132	.132
I									.161	.036	-.474	.562	.029	-.075	-.046	.046
S										-.280	.212	.109	.314	.093	-.033	.033
H											-.719	-.441	-.129	.065	-.032	.032
SL												-.307	.168	-.062	.029	-.029
V													-.040	-.009	.006	-.006
MO														.267	.212	-.212
ER															.479	-.479
CR																-1.000

Table 3: Correlation matrix.

These correlations reveal some redundancy in the features being analyzed. Actually, it is obvious that CR and DR are complementary variables, as far as CR+DR = 1. A high correlation coefficient also shows that one of the variables B or I can be despised without losing too many meaningful information. Therefore, from now on a feature vector of 15 dimensions can be considered.

4.2 Dimensionality analysis

The following analysis consists of finding the dimensions in feature space which best explain the variance of the studied data set. In this way, very low variance dimensions can be detected. This fact implies one of two things: the data can be best expressed using less linear combinations of the original variables than the number of existing variables, or simply there are still some redundant variables. Among the many techniques that can be applied, principal component analysis (PCA) is used in our work. The dimensions obtained by this statistical data reduction technique are linear combinations of the original variables. The weights of each variable are obtained from the eigenvectors of the covariance matrix of the mean-adjusted observed data. The eigenvalues are directly related to the variance associated with the dimensions expressed by
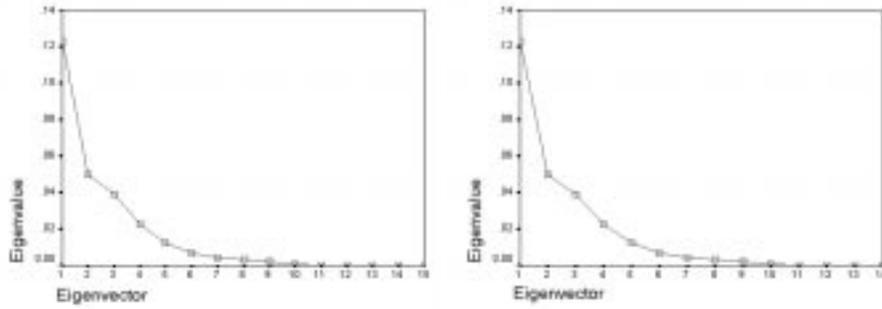
Figure 7: Eigenvalues of the eigenvectors considering ER (left) and not considering it (right).

1st. PC	B	W	R	G	BL	C	Y	M	S	H	SL	V	MO	ER	CR
w. ER	.080	-.182	-.212	.078	.141	.149	.042	-.124	-.011	-.049	.064	-.016	.222	.482	.998
w/o. ER	.080	-.182	-.212	.078	.141	.149	.042	-.124	-.011	-.049	.064	-.016	.222	N/A	.998

Table 4: Weights of each feature in the first principal component considering ER and not considering it.

their associated eigenvectors, so that the eigenvectors with the highest eigenvalues are the dimensions which "best" explain the variance of the data set.

When PCA is based on the covariance matrix, the problem of data scaling can arise. As an example, in our data set, ER has 0 and .05 as minimum and maximum values, while B has .02 and .75 instead, so that the variances in these two dimensions can not be directly compared. As a prove, the eigenvalues spectrum of the PCA is shown in fig. 7 in the case ER is considered and not, as well as the weights of each feature in the first eigenvector, shown in table 4. In both cases, the values obtained are exactly the same.

In this case, the correlation matrix is used instead, resulting in a normalization of the data in order to obtain unitary variances for every feature. The eigenvalues of the eigenvectors obtained in this way are graphically shown in fig. 8. At a quick glance, two superfluous dimensions can be appreciated. Taking advantage of our knowledge about the data acquisition process, a dependence between some of the features, arisen by the fact of considering them as probabilities, was detected. The eight features corresponding to color always sum 1 for every commercial, as well as the ones corresponding to orientations. Therefore, there are 2 redundant variables in the initial representation. This fact is reaffirmed by the plots in fig. 9. If one color and one orientation features are dismissed, all dimensions are somewhat relevant, but if we do not consider two features in any other combination, there will still be irrelevant dimensions.

At this point, having 13 features, PCA still reveals the possibility of one more irrelevant dimension. This is observed due to the fact that 99.99% of the total variance of the original data set is explained by only 12 dimensions. If C is removed from the study, 12 dimensions with significant relevance are obtained, although we think there
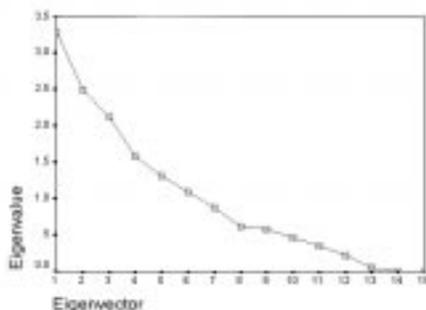
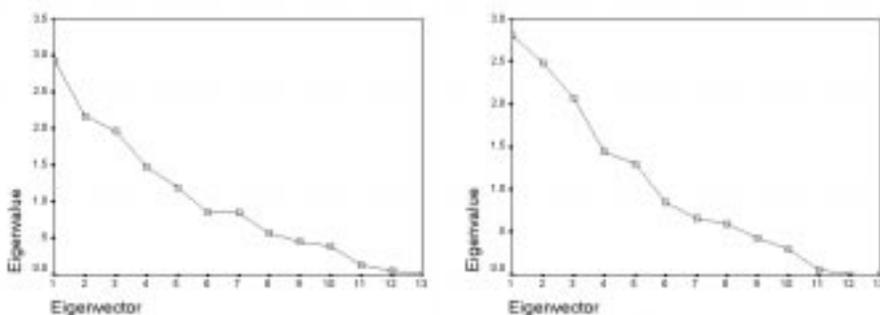Figure 8: Eigenvalues of the eigenvectors of the correlation matrix.

Figure 9: Eigenvalues of the eigenvectors of the correlation matrix neither considering M nor V (left) and neither considering M nor MO, but keeping V (right).

is not enough evidence for doing so and it will be taken into account in the rest of the work.

The final conclusion reached through this part of the work is that the remaining 13 features (B, W, R, G, BL, C, Y, S, H, SL, MO, ER and CR) are meaningful for the representation of the data, although 99.99% of the variance of the data set can be explained by only 12 linear combinations of them. 13 dimensions will be considered in following steps.

4.3   Clusters in feature and semantic spaces

The dimensionality of the original observed data has been reduced to 13 components, which are linear combinations of the initial features computed from the videos. The next step in our study is to analyze the groups of commercials by semantic terms and see if there is a direct relationship with possible groups in the reduced feature space.

Two semantic classifications of the set of commercials under study were obtained

15

Class	Members	Total	Total errors	Error %
Relax	22	70	6	8.57%
Happiness	15	70	2	2.86%
Pleasantness	10	70	5	7.14%
Action	13	70	3	4.29%

Table 5: Results of the classification using a linear discrimination function.

from a number of people: one about the feelings arosen in the viewer by the commercial and another about *consumption values*. The classes proposed were:

Feelings: Action, happiness, relax and stirring.

Consumption values: Utopic, playful, practical and critical.

These classes are not mutually exclusive. Even relax and stirring can be found in one same commercial, probably as a matter of contrasts. In this sense, both semantic spaces can be seen as 4-dimensional spaces, so that a classification would consist of 4 real values. Even so, this can be a problem when it comes to obtaining these values from people. A classification with three degrees of intensity for each value was proposed, corresponding to null, few and much presence in the video. Due to the subjectivity of the classifications, the results obtained from different people were compiled in order to obtain a more reliable one. Regarding to the classification of feelings, the final semantic terms were inferred as combinations of the previous ones and were named relax, happiness, pleasantness (combination of relax and happiness) and action (which can also involve happiness and/or stirring).

Our goal is to find out if commercials are grouped in feature space in the same way they are in semantic space. One way of doing so is trying to find a discriminating function in feature space between the commercials that belong to different semantic classes in a supervised way. Considering this 4 new classes, the simplest discriminating function is the linear one computed for one class against the rest using any well-known algorithm for this purpose. In our case, the perceptron algorithm was used. After that, an automatic classification step tells us about the goodness of the discriminating function, and thus if the commercials are properly grouped in feature space. The results obtained are summarized in table 5. The conclusion is that there exist groups in feature space related with a set of semantic terms which can be well discriminated by simple linear functions. The modelization of the different classes using more complex functions should obviously lead to better results.

On the other hand, the results obtained considering *consumption values* as semantic classes are not so promising. Actually, when taking practical commercials and a linear discrimination function, the perceptron algorithm does not converge to a feasible solution, meaning that such a function does not exist. This is explained by the fact that *consumption values* are excessively high level semantic concepts, which are inferred from the narrative structure of the commercial, more than from its perceptual information. Prior knowledge about the world and much more sophisticated algorithms, like object recognition, facial expression analysis or natural language processing, are required to reach these cognitive levels.

5   Conclusions

This paper has presented a methodology for the preliminary data analysis needed when developing models for representing video information that support both perceptual and semantic level retrieval from visual databases. Our work has been directed to the extraction of visual cues and the analysis of their relevance with respect to semantic concepts. This processes are supported by semiotic studies which suggest what visual features should be used and their possible relationships with the feelings arosen in the viewer and with what is known as *consumption values*.

Visual features about color, orientation, motion and edit rhythm have been extracted and treated in a probabilistic way. Well-known algorithms, like image differentiation, the Hough transform and local color analysis for scene break detection, have been used and enhanced for this purpose. A novel probabilistic approach to color naming has also been implemented and tested.

Features corresponding to 70 commercials have been statistically studied in order to find redundancies in the set of variables. PCA as a dimensionality reduction technique has provided us the dimensions in feature space which best explain the variance of the data set. This analysis has let us reduce the dimensionality of the data from 17 to 13 dimensions, although its total variance can be nearly explained by only 12 linear combinations of the original variables. Finally, the clustering of semantically similar commercials in feature space has been studied considering two different semantic levels: feelings and *consumption values*. This study consisted of trying to discriminate between commercials belonging to a specific semantic class and the rest. A simple linear discrimination function was automatically learnt by means of the perceptron algorithm and its goodness was assessed by classifying the same training data set. In this way, the main conclusion of our work is that a direct relationship exists between visual cues and the semantic level of feelings, i.e. the feelings that arise in the viewer when he/she is watching a commercial are greatly due to the way it is shot, edited and how colors are used. On the other hand, the semantic level of *consumption values* is mainly related to the narrative structure of the commercial, so that visual cues are not so significant in order to recognize different concepts at this level.

Further work is to be done on the semantic characterization of video clips in order to allow their classification at different levels, as well as retrieval tasks. A study about what features are the most relevant in order to discriminate and characterize each semantic class is pending. We are also planning to study the evolution of visual features at shot level, so that contrasts along time can be detected and they can be a good source of information about the semantics in videos.

Acknowledgements

References

[1] D. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13:111–122, 1981.

[2] R. Brunelli, O. Mich, and C. M. Modena. A survey on video indexing. *IRST-Technical Report 9612-06*, 1996.

[3] M. Caliani, C. Colombo, A. Del Bimbo, and P. Pala. Commercial video retrieval by induced semantics. In *Proc. IEEE Intl. Workshop on Content-Based Access of Image and Video Databases*, pages 72–80, Bombay, India, January 1998.

[4] C. Colombo, A. Del Bimbo, and P. Pala. Retrieval of commercials by video semantics. In *Proc. Computer Vision and Pattern Recognition*, pages 572–577, 1998.

[5] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann, 1999.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1977.

[7] B. Furht, S. W. Smoliar, and H. Zhang. *Video and Image Processing in Multimedia Systems*. Kluwer Academic Publisher, Boston, 1996.

[8] C. Kervrann, F. Davoine, P. Perez, R. Forchheimer, and C. Labit. Generalized likelihood ratio-based face detection and extraction of mouth features. *Pattern Recognition Letters*, 18(9):899–912, September 1997.

[9] J. M. Lammens. *A Computational Model of Color Perception and Color Naming*. PhD thesis, State University of New York at Buffalo, June 1994.

[10] C. Papageorgiou, T. Evgeniou, and T. Poggio. A trainable pedestrian detection system. In *Proc. of Intelligent Vehicles*, pages 241–246, Stuttgart, Germany, October 1998.

[11] R. W. Picard. Digital libraries: Meeting place for high-level and low-level vision. In *Proc. Asian Conf. on Computer Vision*, Singapore, December 1995.

[12] R. W. Picard. Light-years from Lena: Video and image libraries of the future. In *Proc. Int. Conf. on Image Processing*, Washington DC, October 1995.

[13] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[14] J. M. Sánchez, X. Binefa, J. Vitrià, and P. Radeva. Local color analysis for scene break detection applied to TV commercials recognition. In *Proc. 3rd Intl. Conf. on Visual Information and Information Systems*, pages 237–244, Amsterdam, The Netherlands, June 1999.

[15] R. J. Schalkoff. *Digital Image Processing and Computer Vision*, pages 213–218. John Wiley, 1989.

[16] B. Takacs and H. Wechsler. Detection of faces and facial landmarks using iconic filter banks. *Pattern Recognition*, 30(10):1623–1636, October 1997.

[17] L. Vilches. *La Lectura de la Imagen. Prensa, Cine y Televisión*. Paidós Comunicación, 1992.

[18] W. Xiong and J. C.-M. Lee. Efficient scene change detection and camera motion annotation for video classification. *Computer Vision and Image Understanding*, 71(2):166–181, August 1998.

[19] J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. In *Proceedings of the ACCV'98*, volume II, pages 687–694, 1998.

[20] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *ACM Conference on Multimedia*, San Francisco, California, November 1995.