Toward a semantic indexing based on images color features similarity?

Emmanuel Veneau[†‡], Julien Pepy[†], Laurent Vinet[†]

[†]*Direction de la Recherche (INA) & [‡]Équipe VISTA (IRISA/INRIA Rennes)*

E-mail: {eveneau,jpepy,lvinet}@ina.fr

published in : International Conference on Image and Signal Processing (ICISP'2001), Agadir, Morocco, May 2001

Considering that similarity of images is an underlying issue in many developed algorithms in the field of automatic video indexing, we wondered about which kind of information is retrieved through the usually extracted color features and associated distances.

In this paper, we tried to understand through CBIR experiments which information one can expect to retrieve. As a first step, we proposed in our work to build a database close to real-world images and to lead experiments on semantic examples in order to evaluate and understand color features and measures.

Key Words: image similarity, color feature, content-based image retrieval, semantic indexing

1. INTRODUCTION

1.1. In search of the semantic content of images

Content-based image retrieval (CBIR) has been a widely studied issue [1, 2, 3], among multimedia database management techniques. Image databases management have applications in various areas, such as medical databases, satellite images, photo-journalism, art, or industry, with a growing demand linked to the development of digital image libraries on Internet. Direct search and query by human users in image databases are based on the *semantic* content of the image, assuming that one is able to extract from an image a compact *semantic* representation of the image through a signature, and to apply on the extracted signature a similarity measure which is perceptually significant [4]. We must admit that it is still an open research area [5, 6].

Beyond the sensitive issue of image retrieval of still images, many other applications to video documents rely on the concept of *similarity*. The main tasks for automatic indexing of video data, such as shot or high-level segmentation, gradual effects detection, or key-frames extraction, use more or less explicitly the concept of *similarity* between frames, and are implemented through features extraction and similarity measures computation [7, 8]. The features extracted to represent an image or a sequence of images through a signature are usually color [9], texture [10], shape [11] or edge descriptors [12], and specific motion descriptors [13] for moving images.

1.2. What is an image made of?

When considering an image (or a sequence of images), three main points of view can be considered :

the semantic nature is what is perceived and understood when watching an image or a video sequence (e.g. *a woman with hat and a sorrowful face* or *French football players cheering after scoring*).

the audio-visual form is how one is induced to perceive and understand an image or a video sequence in a specific way ; this is obviously linked with film direction and editing techniques (e.g. *slow traveling on the scene of crime* or *zoom in, on the singer entering the stage*).

the physical nature of images is the numerical properties of the audio-visual signal available for the algorithms (e.g. for a JPEG image or a MPEG video sequence, the coefficients of each pixel in the $YCrCb$ color space).

On the one hand, when images or video sequences are manually annotated, some *semantic* and *form-linked* data are made available. That is the case for example in the manual annotations attached to video documents archived at INA[1]. On the other hand, algorithms implemented for full-automatic indexing aim at finding out theses data from the *physical* data only.

While doing so, one implicitly assumes that there are, at least, strong links between the various levels of image understanding. According to our experience and discussions with people in charge of manual indexing, it is very often a daring hypothesis. Yet, one can suppose that :

- a given *genre* or a semantic content often implies the use of some specific audio-visual forms (e.g. a rapid edit rhythm for action movies or angle-reverse-angle for dialogues).
- some types of document imply some structural rules (e.g. size of the face in an official portrait or temporal structure of a news program).
- the use of some kinds of techniques can affect the physical content of an image (e.g. a wipe or a close up).

It is nevertheless audacious to suppose that there are somehow regular constants that allow us to retrieve form-linked and semantic data from physical values, and to deduce semantic information from form-linked processes. This issue has been usually tackled mainly through two axes : (i) implementing algorithms in a specific case in order to find valid restricted rules, (ii) trying to extract step by step data of higher semantic levels. However, one can have the feeling that this issue is

[1] for further information, see http://www.ina.fr/Recherche/index.html

often avoided and that results of automatic algorithms suffer from this intrinsic limitation.

Thus, one is led to wonder about the relevance of using algorithms developed with a mathematical point of view and based on *physical* data for *semantic* and *form-linked* indexing tasks. For instance, usual experiments on features extracted and similarity measures proposed often suffer from the gap between the development background and the user's aim, and from the lack of evaluation through data based on actual practice. In particular, for experiments, one can notice that usually one of the following case is encountered :

- features and measures are evaluated on database not numerous and not ambiguous enough;
- image database are built in order to evaluate a specific feature (or a measure) whereas one is confronted to images with various and varying parameters;
- artificial or rebuilt images databases are used with no information on real and complicated image.

If these steps are necessary to design algorithms, one does often not know what to expect from them as on confronts the real world.

1.3. Aims of the present work

In this work, we attempt to have an evaluation of some usual features and similarity measures on a *semantic* and *form-linked* database based on images found in video documents as fictions, news or sport programs. As it is a difficult issue both to construct a coherent semantic database and to evaluate various features and similarity measures on it, we restrict our first experiments to color features. This is justified in the one hand because color features are widely used for automatic indexing tasks and many technics can be found in literature concerning still or moving images, and in the other hand because there is an intuitive link between color and the semantic concept of *set up* or *scene* of the action.

In this contribution, we will present in section 2 a brief review of color features and similarity measures and a description of the retained ones. In section 3, we will described the images database and the experiments conducted. Results obtained will be discussed in section 4, and section 5 will summarize the conclusion of this work.

2. COLOR DESCRIPTORS AND SIMILARITY MEASURES

2.1. Color space

The most widely used color spaces are RGB, HLS, YCbCr, CIE-Lab and Munsell color spaces [14]. For this work, we have selected three of them, namely :

RGB space The color space RGB is an additive colour system, based on the trichromatic components R*ed*, G*reen* and B*lue*, which is very common and easy to implement, but which is not intuitive as far as human visual perception is concerned.

HSL space This system based on H*ue*, S*aturation* and L*ightness* is more intuitive and related to the human visual perception. The mathematical transformation from RGB to HSL space can be found in [14]. The component H appears to be more important than others for visual perception.

YCbCr space This space, which is related to JPEG's image format and MPEG's video format, is based on a component of luminance (Y) and two components of chrominance : B*lue* and R*ed*. This color space was useful to us as we wanted to lead experiments on grey levels only, through the component Y.

2.2. Color signatures

2.2.1. A brief overview

Color features are widely used features [9, 15], and we do not claim to present here an exhaustive review of theses techniques. The basic idea is that colors of pixels in an image are seen as a statistical distribution. Thus, one possible signature is to extract the low-order color moment (i.e. mean, variance, skewness) [15]. The other usual signatures are the color histograms [9] which are the joint probability of the intensities of the three color channels. As histograms are said to be very sparse and, and as a consequence, sensitive to noise, cumulative histograms have been developed [15]. An alternate signature proposed was the extraction of a list of dominant colors or color set [16].

If these global color features are described as relatively robust to small change in content and in image orientation, a drawback often noted is that spatial relations are not considered, which causes many false alarms in automatic indexing systems. Therefore, some authors proposed refinement to distinguish unsimilar images with identical color signatures. Localized or region based descriptors have been proposed for color moments and for histograms [17]. Refinements of histograms have been proposed to take in account more information such as edge density, texturedness, gradient magnitude, or coherence (CCV) [18], or to weight the contribution of a pixel to the histogram through various values such as laplacian, probabilistic or fuzzy measures [19]. Another approach to introduce spatial information is to use color correlogram or auto-correlogram [20].

The following sections will describe more precisely the implemented color signatures.

2.2.2. Color histograms

Considering a three-dimensional[2] color space (x, y, z), quantized on each component to a finite set of colors which correspond to the number of bins N_x, N_y, N_z, the color histogram of the image I is the joint probability of the intensities of the three color channels, that is $\forall 1 \leq i \leq N_x, \forall 1 \leq j \leq N_y, \forall 1 \leq k \leq N_z$: $h(i, j, k) = Card\{p \in I,$ such as $color(p) = (i, j, k)\}$. The color histogram H of image I is then defined as the vector $H(I) = (\ldots, h(i, j, k), \ldots)$.

2.2.3. Region histogram

Region histogram is a weakly spatial constrained histogram, where the image I is partitioned in N_a areas. A color histogram is computed for each area , that is $\forall 1 \leq a \leq N_a, \forall 1 \leq i \leq N_x, \forall 1 \leq j \leq N_y, \forall 1 \leq k \leq N_z$: $h_r(a, i, j, k) = Card\{p \in I,$ such as $color(p) = (i, j, k)$ and $Area(p) = a\}$. The region histogram H_r of image I is then defined as the concatenated vector $H_r(I) = (\ldots, h(a, i, j, k), \ldots)$.

[2] similar results can be deduced when considering only one component, e.g. luminance Y in (Y, C_b, C_r).

2.2.4. Cumulative histogram

A cumulative histogram is computed from a color histogram as follows, $\forall 1 \leq i \leq N_x, \forall 1 \leq j \leq N_y, \forall 1 \leq k \leq N_z : h_c(i,j,k) = \sum_{l \leq i} \sum_{m \leq j} \sum_{n \leq k} h(l,m,n)$. The cumulative histogram H_c of image I is then defined as the vector $H_c(I) = (\ldots, h_c(i,j,k), \ldots)$.

2.2.5. Color coherence vector

Color coherence vector can be seen as a color histogram where pixels in each bin are split between coherent and non coherent pixels [21]. A pixel is said to be coherent if it belongs to a large group of pixels of the same color. We have $\forall 1 \leq i \leq N_x, \forall 1 \leq j \leq N_y, \forall 1 \leq k \leq N_z$:

$$\begin{cases} \alpha(i,j,k) = & Card\{p \in I, \text{ such as } color(p) = (i,j,k) \text{ and } Coh(p) = 1\} \\ \beta(i,j,k) = & Card\{p \in I, \text{ such as } color(p) = (i,j,k) \text{ and } Coh(p) = 0\} \end{cases}$$

The color coherence vector CCV of image I is then defined as the vector $CCV(I) = (\ldots, (\alpha(i,j,k), \beta(i,j,k)), \ldots)$, where $\alpha(i,j,k)$ and $\beta(i,j,k)$ are respectively the coherent and incoherent number of pixels of color (i,j,k).

2.2.6. Auto-correlogram

Auto-correlograms is a simplified version of correlograms based on the probability of two pixels of a given color to be at a given distance in the image I, that is $\forall d \in D, \forall 1 \leq i \leq N_x, \forall 1 \leq j \leq N_y, \forall 1 \leq k \leq N_z : \gamma^{(d)}(i,j,k) = Card\{(p_1,p_2) \in I^2, \text{ such as } color(p_1) = color(p_2) = (i,j,k) \text{ and } dist(p_1,p_2) = d\}$. The auto-corellogram of image I is then defined as the vector $AutoCorr(I) = (\ldots, \gamma^{(d)}(i,j,k), \ldots)$, with D a given set of distances.

2.2.7. Color moments

The moments of the three first orders of the color distribution are computed on image I for each component of the color space. In our work, we consider color histogram as an approximation of the color distribution and compute color moments as follows : $E_x = \frac{1}{N_x} \sum_{i=1}^{N_x} h_x(i)$, $\sigma_x = \frac{1}{N_x} \sqrt{\sum_{i=1}^{N_x} (h_x(i) - E_x)^2}$, $s_x = \frac{1}{N_x} \left(\sum_{i=1}^{N_x} (h_x(i) - E_x)^3 \right)^{\frac{1}{3}}$, where $h_x(i) = \sum_{j,k} h(i,j,k)$. Similar formulas stand for components y and z. The color moments of image I is defined as the nine-dimensional vector $ColMoment(I) = (E_x, E_y, E_z, \sigma_x, \sigma_y, \sigma_z, s_x, s_y, s_z)$.

2.2.8. Dominant colors

We have implemented the algorithm proposed in [22] in order to get a list of colors (x,y,z) associated with a population coefficient p. $DomCol(I) = (\ldots, ((x_i, y_i, z_i), p_i), \ldots)$.

2.2.9. Parameterized signatures

As one can see the color signatures implemented are defined according to various parameters. The tuning of these parameters would deserve a separate study, the values selected for this work, based on the standard values given by the literature are given in table 1.

TABLE 1

Implemented options for color signatures

Color space	Signature	Options implemented	Measures
RGB, HLS	ColorHisto	$N_r = N_g = N_b = 4$ or 8, $N_h = 16$ $N_l = N_s = 8$	$d_1, d_2, d_A, d_{\chi_2}$
RGB, HLS	RegionHisto	$N_r = N_g = N_b = 4$ or 8, $N_h = 16$ $N_l = N_s = 8$	d_1, d_{EMD}
RGB, HLS	CumulHisto	$N_r = N_g = N_b = 4$ or 8, $N_h = 16$ $N_l = N_s = 8$	d_∞
RGB, HLS	CCV	$N_r = N_g = N_b = 4$ or 8, $N_h = 16$ $N_l = N_s = 8$	d_1
RGB, HLS	AutoCorr	$N_r = N_g = N_b = 4$, $N_h = 16$ $N_l = N_s = 8$	d_1
RGB, HLS	ColMoment		d_1
RGB	DomCol	$N_r = N_g = N_b = 4$ or 8	d_{EMD}
Y	LumHisto	$N_y = 16$ or 32 or 64	d_1

2.3. Similarity measures

One can find in literature a huge number of distances, metrics and dissimilarity measures, beginning with the Minkowski's norms up to sophisticated and specific functions. As for color signatures, we just implement standard dissimilarity measures, normalized when possible. The implemented options are presented in table 1.

For color histograms, we have selected the four following dissimilarity measures: L_1 norm [9], which is equivalent to histogram intersection, L_2 norm [9], χ_2-test [8],and L_A based on a quadratic form [23] and a similarity matrix between quantized colors.

Dissimilarity measures selected for region histograms are L_1 norm, and Earth Mover's distance which is described in details in [24].

For cumulative histograms, we considered, according to [8], L_∞ norm, which is equivalent to Kolmogorov-Smirnov test.

To compute dissimilarity over Color Coherence Vector, we have implemented a specific variant of L_1 norm described [21].

For auto-correlogram, an alternative to L_1 norm has been implemented as suggested in [20].

Color moments were associated to a weighted version of L_1 norm [15] and dominant colors to the Earth Mover's distance [24].

3. IMAGE DATABASE DESCRIPTION

As announced in section 1, we intend to build an image database strongly representative of the ordinary video data in order to be able to get results that can be connected with usual automatic indexing algorithms' behaviour.

3.1. Description

Our image database is made of 310 images extracted from video documents on various themes, as follows :

- News program : 67 images presenting studio views and excerpts from three different reports;
- Fiction : 62 images extracted from four excerpts from an episode of *Avengers*;
- Sport program : 19 images of a figure skating broadcast;

- Sport program : 27 images of a rowing broadcast;
- Sport program : 108 images of an athletics meeting;
- Sport program : 14 images of the broadcast presenter;
- Others : 13 various images mainly extracted from credit titles.

Thus, we consider that we have an ambiguous enough database due to the various origin of the images, strongly linked to usual broadcast video document, and with a large variety of lighting condition, framing, shooting angle, focal plane, etc.

3.2. Organization

As explained above, organizing a database according to the concept of image similarity is tricky, though we could lean on various arguments. Two images can be similar because : the same person is presented, the background is identical, the shooting angle is similar, etc.

In order to justify our process and because we are testing global color features, we defined our similarity levels based on the visual colors perception of the images we have. Doing this, we lean on the usual hypothesis that the concept of *set up*, the objects presented, and the color information underlie the studied features. This leads us to build similarity levels defined together on semantic, form-linked and physical considerations.

Level A nearly identical images;

Level B weak variation of objects between the images;

Level C in spite of the changes, a majority of similar objects;

Level D similar objects can be found in the images but they are rare.

As color features are often semantically linked to the concept of *set up*, level A can be associated with the notion of *strong similarity*, level B with the notion of *scene similarity*, level C with the notion of *setting similarity* and level D with the notion of *surrounding similarity*. As far as applications are concerned, level A could correspond to object retrieval task, level B to shot segmentation, level C to high-level segmentation, and level D could be used to construct videos' table of content. Examples of theses similarities are proposed on figure 1. The whole image database was manually labeled.

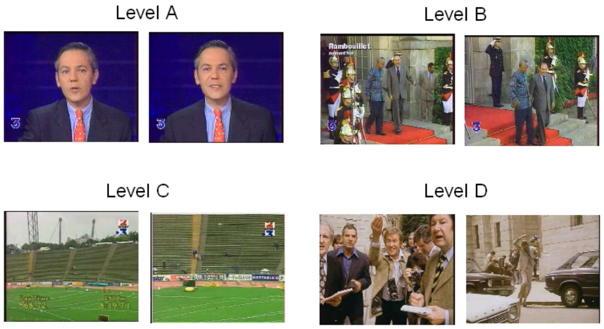

FIG. 1. Examples of similarities between images

4. RESULTS AND COMMENTS

Considering the properties of the database described above, we make experiments in order to understand what kind of information (and particularly *semantic* or form-linked information), the various pairs of options are able to detect. That is why we compared and analyzed the performances of signatures and measures implemented on image queries at different levels of similarity, through our heterogeneous and real-world-like database. Then, in an attempt to classify them, we studied their performances on qualitative tasks based on specific semantic concepts.

4.1. Implemented experiments and evaluation indicator computation

In the first place, we wanted to obtain global results based on defined similarity levels. Considering a given image I_q similar to N other images of the database according to a given level of similarity, if I_q is proposed to the system as a query and only p images out of the N similar images are retrieved in the first N result images ranked by dissimilarity measure, we took the value p/N as a rating of retrieval goodness. The average value was computed for every signatures and measures implemented on all the images at each similarity level. This experiment was rounded off by a qualitative study of specific sets of images for each similarity level.

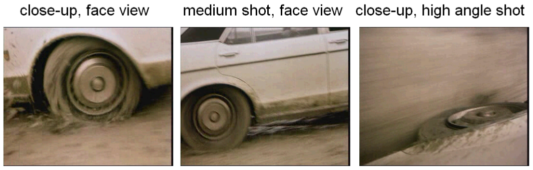

FIG. 2. Example of object similarities with various framing, and shooting angles

Then, we focused on specific, and more qualitative, semantic and/or form-linked retrieval. In particular, we studied the appearance of an object through three images with various framing, and shooting angles (cf. figure 2), and inclusion relationships through eight pairs of images (cf. figure 3). In both cases, the rating considered was the average rank of retrieval. We think that the first example could be a good study of object similarity in the context of a video document, and the second example an important issue related to *set up* and object coherence, or camera movement (zoom in, zoom out).

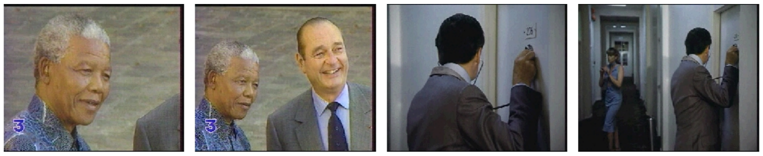

FIG. 3. Two examples of images inclusion

4.2. Results on similarity levels

The results obtained are presented in tables 2 to 4.

4.2.1. About color space effects

Results corroborate the usual belief on color spaces. Signatures on HLS color space, which is supposed to be closer to human visual perception, perform better than on RGB color space. The differences are about 2% for all signatures. When using only the luminance component of YC_bC_r space color, results go down to 10%.

TABLE 2

Results on RGB color space with $N_{bins} = 4$ and $N_{bins} = 8$ for histograms

Options	$N_{bins} = 4$				$N_{bins} = 8$			
	Level A	**Level B**	**Level C**	**Level D**	**Level A**	**Level B**	**Level C**	**Level D**
ColorHisto, d_1	84%	56%	41%	39%	85%	62%	50%	47%
ColorHisto, d_2					81%	55%	42%	41%
ColorHisto, $d_{\chi2}$					85%	63%	50%	47%
ColorHisto, d_A	84%	50%	37%	36%				
RegionHisto, d_1	88%	59%	43%	41%	89%	63%	51%	47%
RegionHisto, d_{EMD}	86%	59%	43%	40%	85%	64%	51%	47%
CumulHisto, d_∞	82%	43%	31%	31%	84%	48%	34%	34%
CCV, d_1	82%	57%	42%	42%	85%	65%	50%	48%
AutoCorr, d_1	84%	59%	50%	53%				
ColMoment, d_1	73%	39%	29%	31%				
DomCol, d_{EMD}	81%	44%	35%	36%	79%	46%	36%	36%

4.2.2. About color quantization

In RGB color space, we obtain better results with a 8 bins per component than with 4 bins, and the improvement is more noticeable for level D than for level A. When the level increases, the similarity between the image in a set decreases, and a few number of bins could account for a higher rate of false positive. Similar results can be observed with luminance only.

TABLE 3

Results on HLS color space

Options	**Level A**	**Level B**	**Level C**	**Level D**
ColorHisto, d_1	85%	66%	51%	49%
ColorHisto, d_2	84%	56%	42%	39%
RegionHisto, d_1	89%	65%	51%	48%
RegionHisto, d_{EMD}	86%	66%	53%	49%
CumulHisto, d_∞	85%	47%	35%	36%
CCV, d_1	85%	65%	51%	49%
AutoCorr, d_1	88%	64%	50%	52%
ColMoment, d_1	78%	39%	32%	35%

4.2.3. About the similarity level

With a few exceptions, retrieval performances of signatures and measures decrease when the level increases. Obviously, this result is due to the heterogeneousness of sets of images of higher level, and to the growing influence of ambiguity between images of different sets. The fact that the sets of higher level contain a wide number of images could explain the few exceptions observed at level D.

TABLE 4

Results on luminance in YC_bC_r color space

Options	Level A	Level B	Level C	Level D
LumHisto, d_1, $N_{bins} = 16$	77%	42%	30%	34%
LumHisto, d_1, $N_{bins} = 32$	76%	44%	30%	34%
LumHisto, d_1, $N_{bins} = 64$	79%	45%	31%	35%

4.2.4. About signatures and measures

Global results prove that six of the signatures and measures are reliable on RGB and HLS : ColorHisto associated with d_1 or d_{χ_2}, RegionHisto associated with d_1 or d_{EMD}, CCV associated with d_1 and Autocorr associated with d_1. Other pairs of options give disappointing results : CumulHisto, ColorHisto with d_2, and especially DomCol and ColMoment.

As expected, RegionHisto performs better than ColorHisto when the similarity is strong between images, and retrieval performances of RegionHisto decrease faster when sets of images become more heterogeneous, even if the d_{EMD} seems more robust. For ColorHisto, the d_{χ_2} seems to be a better measure to be used. From our experiments, AutoCorr seems to be relatively robust to heterogeneousness of the sets of images, which can be a promising research direction.

Some of our results do not match the expected effect. CumulHisto was supposed to be an improvement of ColorHisto, which is not seen through the results. If results with CumulHisto at level A are close to results with ColorHisto, it appears that CumulHisto loose efficiency at higher levels. Results are disappointing too for ColorHisto with measure d_A, both in RGB and HLS color space, as retrieval rates are slightly inferior. This could be explained by the choice of the associated matrix A, which could be mistuned. This is a recurrent issue when dealing with signatures and measures : how to tune their parameters in order to get the best results of each pair of signature and measure. The disappointing results of DomCol could come from the simplified technique [22] used to extract the list of dominant colors. Lastly, according to the experiments we have done, ColMoment seems to be not precise enough when similarity is strong and not adapted when heterogeneousness of images increases.

4.3. Results on specific experiments

Results on specific semantic and/or syntactic retrieval described on section 3 (cf. figure 2 and 3, for example) are presented in tables 5 and 6.

Global results are confirmed by these specific experiments. The pairs of efficient signatures and measures are roughly the same, and HLS seems to be a better color space than RGB.

More precisely, AutoCorr is a signature robust to framing variations and to retrieve the context of an included image, RegionHisto is efficient on both framing

TABLE 5

Results on images with varying framing and shooting angle

Parameter → **Options** ↓	**Framing** (Average rank)	**Shooting angle** (Average rank)	**Framing and shooting angle** (Average rank)
ColorHisto (RGB), d_1	4	6	28
ColorHisto (RGB), d_{χ_2}	6	7	16
RegionHisto (RGB), d_{EMD}	1	4	10
CumulHisto (RGB), d_∞	30	30	51
CCV (RGB), d_1	4	16	22
AutoCorr (RGB), d_1	1	10	6
LumHisto (Y), d_1	35	51	30
ColorHisto (HLS), d_1	1	3	9
AutoCorr (HLS), d_1	1	4	2

and shooting angle variations associated to d_{EMD}, and on inclusion retrieval associated to d_1. CCV is better on framing variations than on shooting angle, and d_{χ_2} leads to decent results but not better than d_1 when associated to ColorHisto.

TABLE 6

Results on included images ($A \subset B$)

image $A \subset$ **image** B **Options**	**Retrieve B from A** (Average rank)	**Retrieve A from B** (Average rank)
ColorHisto (RGB), d_1	2	3
ColorHisto (RGB), d_{χ_2}	2	3
RegionHisto (RGB), d_1	2	4
CumulHisto (RGB), d_∞	11	11
CCV (RGB), d_1	2	3
AutoCorr (RGB), d_1	4	11
LumHisto (Y), d_1	11	19
ColorHisto (HLS), d_1	2	3
CCV (HLS), d_1	2	3

Concerning the nature of the request, retrieving an image B from an image A when $A \subset B$ or similar objects with various framings seems to be quite achievable, whereas the other requests experimented are more complex.

4.4. Remarks about semantic indexing issue

Our approach intends both to evaluate the various pairs of options and to understand if color features and measures could be an answer to tackle some specific semantic indexing task.

Even if some pairs of signatures and measures appeared to us to be more efficient through quantitative results, we could not associate a specific (semantic and/or form-linked) task to a particular pair of options and thus obtain a qualitative description of options.

4.5. Comparison with related works

Finally, we wanted to compare our study with other evaluations of signatures and measures presented in the literature. In [25], the authors compare metrics for the shot boundary detection task, in [26] various metrics are tested for different tasks as classification, image retrieval, unsupervised segmentation. In [25], ColorHisto gave good results and so did CumulHisto. d_1 and d_{χ_2} were quite similar with a slight advantage to d_{χ_2}. In [26], measures d_1, d_2 and d_A were found more or less equivalent, d_{EMD} was a promising measure, d_∞ showed weak results and d_{χ_2} quite better ones. According to our experiments, we can confirm the good results obtained with d_{EMD}, and that d_{χ_2} has a slightly better than Minkovski's metrics.

As in [26], we encounter difficulties to both collect and structure a ground truth where a problem with a known answer can be modeled and where the ambiguity of images can be introduced as a part of the reality of audio-visual documents. We were also concerned with tuning parameters of signatures and measures in order to obtain better results, without resulting in a given methodology.

5. CONCLUDING REMARKS AND FUTURE WORK

In this work, we studied signatures and measures on an images database built from various audio-visual documents and based on four levels of similarity. Moreover, we attempted to analyze the behaviour of these signatures and measures in some cases of semantic and/or syntactic similarity. If we did not obtain a semantic classification of signatures and measures nor a mapping from signatures to level of similarity, we did get some general results on image similarity. From the experiments made, ColorHisto associated to d_1 or d_{χ_2} were proved to be robust and easy to use signatures, CCV appeared as a real improvement on histograms, RegionHisto gave good results mainly for strong similarity. d_{EMD} seems to be a promising measures that can be efficiently used associated to RegionHisto. Autocorr seems to be a quite interesting signature, robust to spatial changes in the image.

Moreover, we have proposed the first step to set an evaluation methodology on various features and measures for qualitative tasks.

Future research directions will be to build a more precise ground truth, to try to tackle the difficult issue of tuning parameters linked to signatures and measures, and to achieve more semantic information through multi-criteria features or measures.

REFERENCES

1. M. De Marsicoi, L. Cinque, and S. Levialdi, "Indexing pictorial documents by their content : a survey of current techniques," *Image and Vision Computing*, vol. 15, pp. 119–141, 1997.
2. R. M. Bolle, B.-L. Yeo, and M. M. Yeung, "Video query : Research directions," *IBM Journal of Research and Development*, vol. 42, no. 2, pp. 233–252, 1998.
3. Yong Rui, Thomas S. Huang, and Shih-Fu Chang, "Image retrieval: Current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 39–62, 1999.
4. A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook : Content-based manipulation of image databases," in *Storage and Retrieval for Image and Video Databases II*, SPIE, Ed., San Jose, CA., 1994, vol. 2185.
5. Navenka Dimitrova, "The myth of semantic video retrieval," *ACM Computing Surveys*, vol. 27, no. 4, 1995.
6. J. Huang, S. R. Kumar, and R. Zabih, "An automatic hierarchical image classification scheme," in *Sixth ACM Intl. Multimedia Conf.*, 1998, pp. 219–228.

7. F. Idris and S. Panchanathan, “Review of image and video indexing techniques,” *Journal of Visual Communication and Image Representation*, vol. 8, no. 2, pp. 146–166, 1997.

8. R. Brunelli, O. Mich, and C.M. Modena, “A survey on the automatic indexing of video data,” *Journal of Visual Communication and Image Representation (JVCIR)*, vol. 10, pp. 78–112, 1999.

9. Michael J. Swain and Dana H. Ballard, “Color indexing,” *Int. Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

10. B.S. Manjunath and W.Y. Ma, “Texture features for browsing and retrieval of image data,” *Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 836–842, 1996.

11. B. Scassellati, S. Alexopoulos, and M. Flickner, “Retrieving images by 2d shape: a comparison of computation methods with human perceptual judgments,” in *Storage and Retrieval for Image and Video Databases*, San Jose, CA, 1994, vol. 2185, pp. 2–14.

12. R. Lienhart, W. Effelsberg, and R. Jain, “Towards a visual grep : A systematic analysis of various methods to compare video sequences,” in *Storage and Retrieval for Image and Video Databases VI*, K. Sethi and Ramesh C. Jain, Eds., Ishwar, 1998, pp. 271–282.

13. Ronan Fablet, Patrick Bouthemy, and Patrick Perez, “Statistical motion-based video indexing and retrieval,” in *Conf. Int. sur la Recherche d'Informations Assiste par Ordinateur*, Paris, France, 2000.

14. Adrian Ford and Alan Roberts, “Colour space conversions,” Technical, University of Westminster, August 11 1998.

15. Markus Stricker and Markus Orengo, “Similarity of color images,” in *Storage and Retrieval for Image and Video Databases III*, SPIE Proceedings Series, Ed., 1995, vol. 2420, pp. 318–392.

16. John R. Smith and Shih-Fu Chang, “Tools and techniques for color image retrieval,” in *Storage & Retrieval for Image and Video Databases IV*, Sethi I. K. and Jain R. C., Eds., San Jose, CA, 1996, vol. 2670, pp. 426–437.

17. Markus Stricker and Alexander Dimai, “Color indexing with weak spatial constraints,” in *Storage and Retrieval for Image and Video Database IV*, SPIE, Ed., San Jose, CA, 1996, vol. 2570.

18. Greg Pass and Ramin Zabih, “Comparing images using joint histograms,” *Journal of Multimedia Systems*, 1998.

19. C. Vertan and N. Boujemaa, “Upgrading color distributions for image retrieval: can we do better ?,” in *International Conference on Visual Information Systems*, Lyon, France, 2000.

20. J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, “Image indexing using color correlograms,” in *Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997, pp. 762–768.

21. Greg Pass, Ramin Zabih, and Justin Miller, “Comparing images using color coherence vectors,” in *ACM Multimedia*, Boston, USA, 1996, pp. 65–73.

22. Ing-Sheen Hsieh and Kuo-Chin Fan, “An adaptative clustering algorithm for color quantization,” *Pattern Recognition Letters*, vol. 21, pp. 337–346, 2000.

23. James Hafner, Harpreet S. Sawhney, Will Equitz, Myron Flickner, and Wayne Niblack, “Efficient color histogram indexing for quadratic form distance function,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729–736, 1995.

24. Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, “The earth mover's distance as a metric for image retrieval,” Technical Report STAN-CS-TN-98-86, Department of Computer Science, Stanford University, September 1998.

25. Ralph M. Ford, Craig Robson, Daniel Temple, and Michael Gerlach, “Metrics for shot boundary detection in digital video sequences,” *Multimedia Systems*, vol. 8, pp. 37–46, 2000.

26. Jan Puzicha, Joachim M. Buhmann, Y. Rubner, and C. Tomasi, “Empirical evaluation of dissimilarity measures for color and texture,” in *ICCV*, 1999, vol. II, pp. 1165–1172.

Acknowledgements Images, part of the AIM corpus, were reproduced thanks to INA, Department Research and Experiments. We particularly wish to thank Michèle Wautelet for her invaluable help and advice.