Automatic difference measure between movies using dissimilarity measure fusion and rank correlation coefficients

Nicolas Voiron Alexandre Benoit Patrick Lambert
LISTIC - Université de Savoie - BP 80439 - 74940, Annecy le Vieux, France
{ Nicolas.Voiron, Alexandre.Benoit, Patrick.Lambert} @univ-savoie.fr

Abstract

When considering multimedia database growth, one current challenging issue is to design accurate navigation tools. End user basic needs, such as exploration, similarity search and favorite suggestions, lead to investigate how to find semantically resembling media. One way is to build numerous continuous dissimilarity measures from low-level image features. In parallel, an other way is to build discrete dissimilarities from textual information which may be available with video sequences. However, how such different measures should be selected as relevant and be fused ? To this aim, the purpose of this paper is to compare all those various dissimilarities and to propose a suitable ranking fusion method for several dissimilarities. Subjective tests with human observers on the CITIA animation movie database have been carried out to validate the model.

1 Introduction

In many domains, the amount of multimedia dataset increases dramatically. Whatever may be the content, we need efficient ways to search and navigate in this data. In addition, some more exploration functionalities are required. For example, favorites and suggestions can help end-user to focus on its own area of interest. All this leads to investigate the measures of similarities between multimedia objects, concepts and contents.

Many methods has been explored to measure the similarity between still images [4]. For videos, the amount of data is much larger, so that the analysis complexity of the low level information increases even more. However, descriptors extracted from images or videos are much lower level than the interpretation level that end-user expects. This is called the semantic gap.

In this field, challenges such as TRECVID [8] with tasks such as Semantic INdexing, Content-based Copy Detection, INstance Search... or MediaEval[1] with automatic genre classification are a good mirror of such research activity domains. All these tasks are generally based on similarity or dissimilarity measures.The approach proposed in this paper is slightly different, since our aim is to build a natural dissimilarity closer to human opinion. In TRECVID tasks such as copy detection or instance search, the similarities are measured on short parts of videos. In our case, dissimilarity measure are considered on the whole video. Also, in automatic classification, a lot of approaches are dedicated to genre or sub-genre classification. Using similar data or features, our work is different as we are looking for a dissimilarity measure between video pairs. Such pairs may be part of the same class or eventually from different classes.

In this paper, we work with the dataset used by Benoit & al. in [2]. It is composed of short animated movies from the CITIA[2] database which is a part of all the movies shown in the yearly "International Animated Film Festival", which takes place in Annecy (France). In this database, 51 representative movies were selected and a human annotation has been performed. Those annotations, described more precisely in section 2, provide a five class clustering of the 1275 dissimilarities between all the movie pairs. Added to this initial data which will constitute our ground truth, other characteristics are available. These characteristics are of two type depending on their origin. Some are extracted from images (color histogram, ...), the second type being composed of textual information associated to each movie (original title, ...). More details about these characteristics will be given in the following.

In this paper, deepening Benoit & al. [2] work which was exclusively focused on low level image features, we investigate the addition of the textual information. We quantify the improvement provided by this data source and propose a fusion strategy using both text and image to produce a better similarity measure. The originality of this paper is the use of the rank correlation coefficients with the movie's informa-

[1] http://www.multimediaeval.org
[2] http://www.citia.info

Original title	Year	Duration	Country
Casa	2003	07:07	France
Circuit marine	2003	07:50	France Canada
David	1977	08:45	Netherlands
Gazoon	1998	03:30	France

Audience	Genre
12-15_years Young_adults Adults	Artistic Dramatic
All_publics	Adventure
All_publics	Funny
All_publics	Artistic

Table 1. Textual data sample

tion. The originality is also in the proposal of the successive sorting method to produce the fused measure. The final aim is to get an automatic measure of how a movie is similar to an other one, this measure being as close as possible to human opinion.

This paper is organized as follows. In section 2, we describe how those dissimilarity measures are designed. Since human annotations and textual data produce discrete dissimilarities while image data give continuous ones, managing all these similarity values cannot be performed directly. So, we use a rank comparison approach. In this way, different rank correlation coefficients are defined and an enhanced dissimilarity fusion is proposed in section 3. Comparative results between different dissimilarities and the validation of the fusion method are presented in section 4.

2 Dissimilarity measures

As Batagelj [1] describes it, a dissimilarity measure d on a set E is a function from $E \times E$ to $\mathbb{R}^+$ which is symmetric and satisfies $d(x, x) = 0$. If $d(x, y) = 0$ implies $x = y$, then the dissimilarity is said to be proper. Moreover, if the triangular inequality ($\forall (x, y, z) \in E^3 \;\; d(x, y) + d(y, z) \leq d(x, z)$) is satisfied, the dissimilarity is a distance function (or metric). In this paper, we work with the following listed dissimilarity measures always normalized within range $[0, 1]$:

(i) The first dissimilarity is the *human annotation*. Three human observers performed a manual annotation as described in [2]. They assigned a similarity degree chosen within a 5 point rating scale (very similar, similar, median, different, very different) to the 1275 different pairs of the 51 animated movies from CITIA. This constitutes a 51×51 symmetrical matrix (diagonal is 0) which is recoded to five different regular increasing values (from 0.1 associated to very similar, to 0.9 associated to very different), 0 being voluntarily excluded to satisfy the proper criteria. For each movie pair, we retained the median value of the three human observers. By selecting always the most consensual value, the median provides a discrete dissimilarity with the same 5 different values.

Movie couple	d_{year}	d_{dur}	d_{ctry}	d_{gnr}
Casa / Circuit marine	0	0.036	0.5	0
Casa / David	0.52	0.082	1	0
Casa / Gazoon	0.1	0.181	0	0.5
Circuit marine / David	0.52	0.046	1	0
Circuit marine / Gazoon	0.1	0.217	0.5	0
David / Gazoon	0.42	0.263	1	0

Table 2. Textual dissimilarities sample

(ii) The second dissimilarity family is composed of the normalized Euclidean distances between color and rhythm features proposed in [2]. In the following we only use two aggregated measures of these dissimilarities. The first one is a weighted average. The second one is based on the Choquet integral [5] which considers the interactions between the different feature dissimilarities in addition to the weighted average. In both cases, the weights are adjusted thanks to a learning step using the human annotation. See [2] for more details.

(iii) The third family is the *textual dissimilarities*. An example of textual information associated to four movies is presented in table 1.

(a) All the movies have been produced in the last 50 years. So, in equation (1), we propose a normalized dissimilarity measure based on the "Year" information where $year(x)$ is the year of the movie x release date.

$$d_{year}(x, y) = \frac{|year(x) - year(y)|}{50} \quad (1)$$

Table 2 provides in column 2, this dissimilarity for the four movies given in table 1.

(b) Most of the movies are less than 20 minutes long. Then, in equation (2), we propose a normalized dissimilarity measure between 2 movies x and y based on the "Duration" information.

$$d_{dur}(x, y) = \min\left(1\,,\; \frac{|dur(x) - dur(y)|}{1200}\right) \quad (2)$$

Table 2 provides in column 3, this dissimilarity for the four example movies given in table 1.

(c) For a movie x, all other textual data is described by a set E_x which is a list of words or keywords. For example, "drawing on cells" and "drawing

on paper" are considered as two different keywords.

In equation (3), we propose the classical normalized dissimilarity measure derivated from the Jaccard index [3].

$$d_{.}(x,y) = 1 - \frac{|E_x \cap E_y|}{|E_x \cup E_y|} \quad (3)$$

Table 2 provides in column 4 and 5, this dissimilarity based on the "Country" and "Genre" criteria for the same movies.

In the particular case of the synopsis, we follow the same approach, but we first convert all the words to their lemmas/roots by using a lemmatization software[3]. Punctuation and repeated words are skipped. A sample of synopsis lemmatization is "facetious bird torment ostrich help friend elephant"

(iv) The fourth family is composed of two *low reference dissimilarities* obtained by a random uniform distribution. The first is continuous while the second is discrete.

3 Rank correlation and successive sorting

With these dissimilarities, two questions can be asked:

- Are these objective dissimilarity measures close to the human perceptual dissimilarity measures ?
- How to fuse these dissimilarities to get a global measure closer to the human dissimilarity measures ?

To answer these questions, we need a way to compare and to merge dissimilarities. Numerical approaches such as mean square error, dissimilarity average, ... fail because the range of the used dissimilarities may be different. So, to overcome this difficulty, we propose to use rank correlation.

3.1 Kendall's tau

As in [6], lets us consider object triples (in our case objects are movies). Considering an n object set $E = \{x_1, ..., x_n\}$ and two dissimilarity measures d_1, d_2, we consider that an object triple (x_i, x_j, x_k) is:

(i) *concordant* if:

$$\begin{cases} d_1(x_i,x_j) < d_1(x_i,x_k) \\ d_2(x_i,x_j) < d_2(x_i,x_k) \end{cases} \text{ or } \begin{cases} d_1(x_i,x_j) > d_1(x_i,x_k) \\ d_2(x_i,x_j) > d_2(x_i,x_k) \end{cases}$$

[3]http://www.sphinx-soft.com

(ii) *discordant* if:

$$\begin{cases} d_1(x_i,x_j) > d_1(x_i,x_k) \\ d_2(x_i,x_j) < d_2(x_i,x_k) \end{cases} \text{ or } \begin{cases} d_1(x_i,x_j) < d_1(x_i,x_k) \\ d_2(x_i,x_j) > d_2(x_i,x_k) \end{cases}$$

(iii) *tied* (neither concordant nor discordant) if:

$$d_1(x_i,x_j) = d_1(x_i,x_k) \;\; or \;\; d_2(x_i,x_j) = d_2(x_i,x_k)$$

The *Kendall's tau* coefficient on object triples is:

$$\tau = (C_3 - D_3)/N_3 \quad (4)$$

where C_3 and D_3 are the number of concordant and discordant triples amongst the N_3 considered triples. Kendall's tau takes its values between -1 and 1. A zero coefficient means that the two dissimilarities are independent. A coefficient equal to 1 (respectively -1) means that the dissimilarity rankings are the same (respectively opposite). More generally, the closer to 1 the index, the better the dissimilarity agreement.

3.2 Goodman-Kruskal's gamma and discreteness index

However, if at least one of the two dissimilarities is strongly discrete, the Kendall's tau could be positive and close to zero whereas concordances highly outnumber discordances. The reason is that the number of tied triplets is very much larger than the number of concordances or discordances. So, for discrete dissimilarities, we need an other index indifferent to tied triples. We propose to use the *Goodman-Kruskal's gamma* described by Podani [7]:

$$\gamma = (C_3 - D_3)/(C_3 + D_3) \quad (5)$$

For continuous dissimilarities, it is equal to Kendall's tau. Its only difference behaviour is to be insensitive to tied triples. With those two previous indexes, we deduce a third one, which is an index quantifying globally discreteness of the two dissimilarities, the *percentage of untied triples*:

$$\pi = \tau/\gamma = (C_3 + D_3)/N_3 \quad (6)$$

With those previous indexes, we can efficiently compare all various dissimilarities. In addition, using the Goodman-Kruskal's gamma with the human annotation, we can quantify the quality of the partial ranking produced by discrete dissimilarities.

3.3 Successive sorting fusion

The problem is to build a dissimilarity which combines all the different dissimilarities, knowing that some are discrete and some are continuous. The solution we propose is to use a lexicographic successive sorting approach.

Movie couple	d_{ctry}	d_{year}	d_{dur}	rk	d_f
Casa / Gazoon	0	.	.	1	0.167
Casa / Circ. m.	0.5	0	.	2	0.333
Circ. m. / Gazoon	0.5	0.1	.	3	0.5
David / Gazoon	1	0.42	.	4	0.667
Circ. m. / David	1	0.52	0.046	5	0.833
Casa / David	1	0.52	0.082	6	1

Table 3. Successive sorting sample

Given two ordered sets A and B, the lexicographical order on the Cartesian product $A \times B$ is defined as $(a, b) \leq (a', b')$ if and only if $a < a'$ *or* $((a = a')$ *and* $b \leq b')$

The first step of this approach is to identify a hierarchy, or order, between the dissimilarities (acting as the letter hierarchy in an alphabet). The way we define this hierarchy will be detailed at the end of this section. It can be noted that, as we use a lexicographic order, the successive sorting will be stopped as soon as a continuous dissimilarity will be used in the hierarchy (no tied couples with a continuous dissimilarity). After the identification of a specific order between the dissimilarities, we rank movie couples according to this lexicographic order. Finally, a fused normalized dissimilarity could be for instance obtained by dividing all ranks by the total number of pairs.

An example of this process is given in table 3 with the 6 pairs of the 4 movies shown in table 1 and 2. An arbitrary sorting order has been chosen: "Country", "Year" and "Duration". rk is the rank obtained by applying the lexicographic successive order. In table 3, the first of the 6 couples is the only one with $d_{ctry} = 0$. Its rank is 1. Next, second and third couples have both $d_{ctry} = 0.5$. They are separated by the second sorting criteria d_{year}. So, their ranks are respectively 2 and 3, according to the corresponding d_{year} values order, and so on. Finally, the global dissimilarity d_f is: $d_f = 1/6$ for the first couple, $d_f = 2/6$ for the second, etc.

A related issue is the choice of the dissimilarity order for the successive sorting. We propose a method based on the Goodman-Kruskal's gamma called the *remaining gamma*. Let d_{hu} be the human annotation and d_1, ..., d_p the p other dissimilarity measures. At first, all the p Goodman-Kruskal's gamma between d_{hu} and d_i are computed. The highest gamma is selected and its associated dissimilarity $d_{(1)}$ is put at the top of the hierarchy and is used to rank all the possible pairs. Next, we compute the $p - 1$ remaining gamma which are Goodman-Kruskal's gamma between d_{hu} and d_i considering only still tied pairs which have not been ordered with dissimilarity $d_{(1)}$. This approach is applied until there is no more dissimilarity or tied pairs.

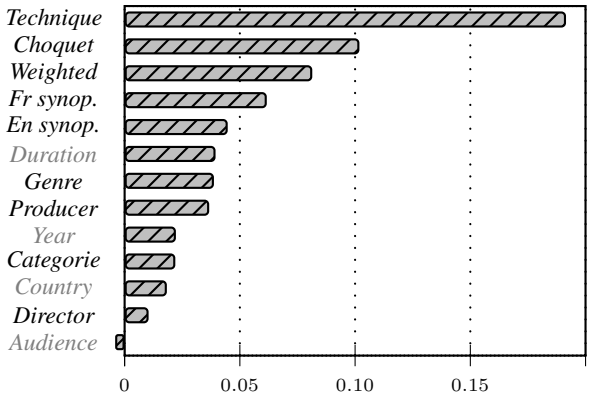

Figure 1. Kendall's tau

4 Results

4.1 Descriptions and first comparisons on individual dissimilarities

Figure 1 shows the Kendall's tau on object triples between human annotation and the thirteen automatic dissimilarity measures on the 51 considered movies. All the eleven textual dissimilarities are presented. The two others, denoted "Choquet" and "Weighted", come from image features (see section 2(ii)).

The most striking point is that "Technique" is the best and well ahead of all the other dissimilarity measures. So, if only one dissimilarity has to be chosen to simulate our human behaviour, "Technique" is the most appropriated. The next best performing features are the two low-level image based dissimilarities. Kendall's tau corroborate Benoit & al's comparison: Choquet fusion is better than the weighted sum. All the remaining features are textual dissimilarities and perform less than the image based features. This can be explained by the discreteness of the textual dissimilarities and the continuity of the others. However, those dissimilarities are not necessarily out of interest. Then, to exclude the tied values, we need to look at the second index.

Figure 2 displays the Goodman-Kruskal's gamma on object triples between human dissimilarity and the same thirteen dissimilarity measures as those presented in figure 1. At first we notice that "Technique" is still the best one. Knowing $\gamma_{Technique} \approx 0.716$ and solving equation (5), we find $C_3 = 6.04 \times D_3$ which means that concordances with human opinion are more than six times larger than the discordances. Knowing $\tau_{Technique} \approx 0.192$, with equation (6) we obtain $\pi_{Technique} \approx 27\%$, which means that 27% of the movie pairs have been ordered by $\gamma_{Technique}$ (23% of concordances and 4% of discordances). In comparison, for the Choquet's dissimilarity, $\tau_{Choquet} \approx 0.102$, $\gamma_{Choquet} \approx 0.162$, $\pi_{Choquet} \approx 63\%$ and concordances are 1.4 times larger than the discordances. As Choquet's dissimilarity is continuous, it consists of distinct values. Then

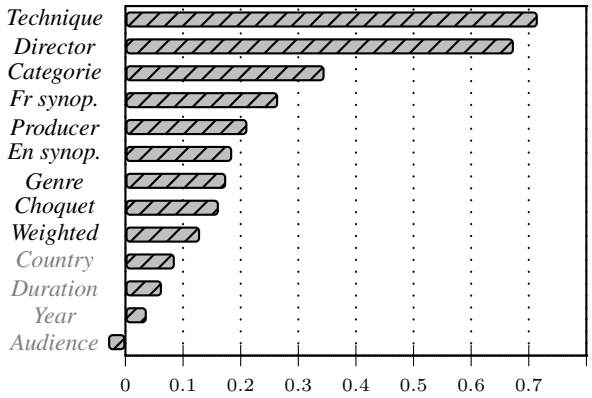

Figure 2. Goodman-Kruskal's gamma

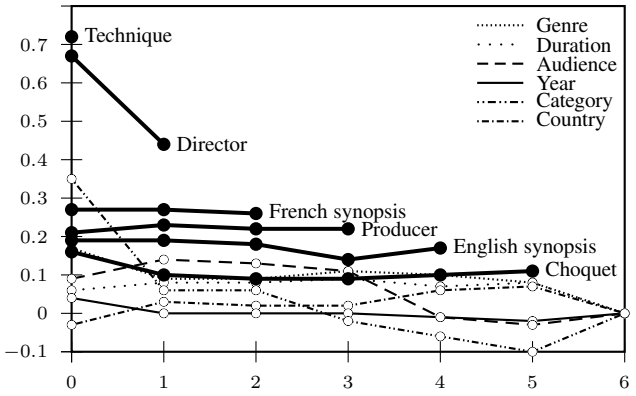

Figure 3. Remaining gamma across successive sorting steps (bold lines = used dissimilarities, thin lines = unused)

object triples cannot be tied by the Choquet's dissimilarity but rather by the human annotation dissimilarity. Consequently, 63% is the percentage of object triples untied by the human annotations. Then, the difference from 27% (using only "Technique") to 63% provides a room for progress of 36%. In other words, it means that among these 36%, it is possible to find other additional criteria different from "Technique" which could increase its 23% concordances when used alone.

An other point is that $\gamma_{Director}$ is almost close to $\gamma_{Technique}$ but $\tau_{Director}$ is very low. This indicates that "Director" dissimilarity is a good criteria for ranking. However, due to the numerous tied values, it ranks only few pairs. In the used database, for each movie, there is only one director or sometimes two co-directors who are almost always different. This means that human observers classify movies from the same director more similar than from other movie directors. Thanks to the fact that directors frequently use the same techniques in their different movies, a dependency between "Director" and "Technique" could exist. In the section 4.2, the proposed remaining gamma method will show if the Director's concordances are totally included, partially included or not included in the Technique's concordances.

In the same way, the Goodman-Kruskal's gamma for "Category", "Producer", "Genre", "French and English synopsis" dissimilarities are greater than the low-level image ones. And the same dependencies between them could appear. "French and English synopsis" could be linked. Same for "Producer" and "Director", "Synopsis" and "Genre"...

Finally, in order to better identify the significance of the less performing features ("Country", "Duration", "Year" and "Audience"), we compare them against two random variables, one continuous and one discrete, each within range $[0; 1]$. Kendall's tau and Goodman-Kruskal's gamma have been computed for 10,000 samples. On the obtained Gaussian distributions, less than 0.5% of the random dissimilarities have a Kendall's tau out of the range $[-0.05; 0.05]$ and a Goodman-Kruskal's gamma out of the range $[-0.1; 0.1]$. As a consequence, figures 1 and 2 present in gray out, four dissimilarity measures, which remain in the random ranges. Particularly, "Audience" indexes are near zero negative. This criteria is a subjective information that may be composed of non-homogeneous entries provided by different human observers. So "Audience" is not relevant for our work. "Duration" and "Year" are objective information, but apparently not usable (at least for the considered set of movies).

4.2 Final fusion

The successive sorting, described in section 3, is suited to obtain a fused dissimilarity improving the Kendall's Tau for several discrete dissimilarities. A continuous dissimilarity could be considered to end the successive sorting. The Goodman-Kruskal's gamma is dedicated to select the sorting order. The first idea could be to select features by descending gamma order obtained from figure 2. But as observed in section 4.1, the dissimilarities can be linked. So we propose to use the remaining gamma technique described in section 3.

Figure 3 shows the remaining gamma through the different steps of the successive sorting. In the proposed situation, six steps are necessary to converge. This diagram aggregates all the method process. Initially, at the zero step, the remaining gammas represent exactly the Goodman-Kruskal's gammas of each dissimilarity; "Technique" gives the best value. Then, all pairs are sorted using "Technique" dissimilarity values. The gamma is computed on the remaining tied object triples in order to identify the most significant criteria to use at the next sorting step. The result is visible at the abscissa 1: the second best result is "Director". This is repeated until step 6 when all pairs are sorted (thanks to the use of "Choquet" which is a continuous dissimilarity) with consequently a zero remaining gamma.

First visible result is that "Director" gamma strongly decreases at step 1. This seems to confirm than "Director" is partially linked to "Technique". However, "Director" is always the second sorting criteria. It can also be noted that "Category" strongly decreases and is no more relevant. This observation demonstrates the interest of the remaining gamma method to exclude redundant information. Next, "French synopsis", "Producer", "English synopsis" and "Choquet" similarity are successively used. With its distinct continuous values, "Choquet" necessarily ends the process. "Category" and five other dissimilarities shown with the thin lines are not taken into account.

Finally, compared to the "Technique" criteria used alone, this fusion process enhances results as the following: first, tied values amount is strongly decreased from 73% to 37%. Second, concordance with human opinion is improved significantly from 23% to 44.5% ($\tau_{rg} \approx 0.260$ and $\gamma_{rg} \approx 0.414$). However discordances are increased in a lower way from 4% to 18.5%. As a conclusion, the proposed fusion allows nearly half of the database to be automatically sorted as human would do.

From a computational point of view, as shown in figure 3, 12 dissimilarities have been used. In the worst case, during the iterative fusion, a maximum of 78 gammas should have been computed. However, in our 6 step context, only 63 were needed. Compared to a rough approach, finding the best performing sorting would lead to analyse all the 479,001,600 possible combinations. After experiment, in our context, such rough approach gives 57 sorting performing better than our remaining gamma method result. However, the highest value ($\tau_{best} \approx 0.263$) only increases performance by one percent above our method. This highlights the remaining gamma method interest in term of "quality" versus "computation time".

4.3 Method evaluation using cross validation

To evaluate this fusion method, the training set is separated from the validation set. 34 movies are randomly taken from the 51 movies dataset for training while the remaining 17 are used for testing. The lexicographic order is obtained on the training set using the remaining gamma method. Fused dissimilarity measure is computed. Kendall's tau are computed on both dataset for results comparison.

To obtain significant results, we applied this operation 1000 times with different randomly subsets. The average of the kendall's tau computed on the learning sets is about 0.294 with a standard deviation about 2.1%. On the validation set values are 0.268 and 5.0%. On this validation set, the obtained Kendall's tau is close (91%) to the value obtained on the learning set.

5 Conclusion

In this paper, we proposed a new automatic dissimilarity measure between movies reproducing human opinion. Such a measure can be used to help user in selecting movies he likes, and avoid movies he does not appreciate. This solution is based on a fusion between image and textual dissimilarities. As the aggregation of such different information cannot be achieved in a numerical way, we have proposed an original solution based on correlation ranks. Performances show that concordances with human opinions are improved by using fusion (from 23% using the best single criteria to 44.5% with fusion). The evaluation described in 4.3 validates this fusion method.

Future work can extend the present approach in many ways. When thinking about a potential application, one can imagine a system relying on a global ranking such as the fused dissimilarity discussed in the paper. It would allow a user to query the system with its own movie and retrieve all the resembling media within a database. An action is also initiated to deploy a collaborative survey software for enlarging the number of movies humanly annotated and also the number of annotations. Another way of improvement could be done on "Technique" or "Genre" dissimilarities by using ontologies rather than the Jaccard index which does not operate on semantic dimension of this data. Similarly, for synopsis, semantic networks could be used instead of cardinal index.

References

[1] V. Batagelj and M. Bren. Comparing resemblance measures. *Journal of classification*, 12:73–90, 1995.

[2] A. Benoit, M. Ciobotaru, P. Lambert, and B. Ionescu. Similarity measurement for animation movies. *MMM (1)*, pages 350–358, 2011.

[3] F. Brucker and J.-P. Barthélemy. *Éléments de classification*. Hermes, London, 2007.

[4] R. Datta, D. Joshi, J. Li, and S. Z. Wang. Image retrieval : Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40, 2008.

[5] M. Grabisch, I. Kojadinovic, and P. Meyer. A review of methods for capacity identification in choquet integral based multi-attribute utility theory : Applications of the kappalab r package. *European Journal of Operational Research*, 186:766–785, 2008.

[6] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[7] J. Podani. A measure of discordance for partially ranked data when presence/absence is also meaningful. *Coenoses*, 12:127–130, 1997.

[8] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, USA, 2006. ACM Press.