

Diversity Ranking for Video Retrieval from a Broadcaster Archive

Xavier Giro-i-Nieto, Monica Alfaro, and Ferran Marques
Technical University of Catalonia (UPC), Barcelona, Catalonia / Spain
{xavier.giro, ferran.marques}@upc.edu

ABSTRACT

Video retrieval through text queries is a very common practice in broadcaster archives. The query keywords are compared to the metadata labels that documentalists have previously associated to the video assets. This paper focuses on a ranking strategy to obtain more relevant keyframes among the top hits of the results ranked lists but, at the same time, keeping a diversity of video assets. Previous solutions based on a random walk over a visual similarity graph have been modified to increase the asset diversity by filtering the edges between keyframes depending on their asset. The random walk algorithm is applied separately for every visual feature to avoid any normalization issue between visual similarity metrics. Finally, this work evaluates performance with two separate metrics: the relevance is measured by the Average Precision and the diversity is assessed by the Average Diversity, a new metric presented in this work.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: [Retrieval models]; I.4.10 [Image Processing and Computer Vision]: Image Representation multidimensional

General Terms

Algorithms, Measurement

Keywords

Image ranking, Video retrieval, Similarity graph, Average diversity

1. MOTIVATION

Television broadcasters store nowadays large and growing amounts of video data in their local archives that require efficient retrieval techniques. The traditional text-based queries and indexing strategies are being complemented with new descriptors automatically extracted from the visual and audio content. These new type of features open the door to

promising possibilities in the video indexing and retrieval domains, such as query by example or high-level concept detectors.

Most video retrieval systems present the search results as a set of keyframes displayed on a graphical user interface. These keyframes are automatically generated from the video data at ingest time by an image processing algorithm. Keyframes become the representation units of the video assets as the user can obtain relevant information about the video by looking at its related keyframes. Keyframes offer many advantages as a representation unit when compared to textual metadata because humans can quickly interpret them and the amount of information per pixel they contain is normally far over the one coded in text.

The documentalists that work in a broadcaster archive usually produce textual metadata at the video asset scale, providing little or none information at the keyframe level. For example, in an interview to a popular character, the textual metadata of the video asset will typically contain the name of this person, although the asset may also contain keyframes where only the interviewer appears. As a result, the manual annotation of the video asset will not apply to all keyframes and, in extension, not all keyframes retrieved by a text-based search will be relevant for the query.

The mismatch between the temporal resolution of the video annotations (several seconds) and the temporal resolution represented by the keyframes (a precise moment) generates a problem when choosing which keyframes are to be shown to the user. One extreme option would be to show all keyframes for every video asset matching the query. As discussed in the previous paragraph, this solution would probably produce the selection of several irrelevant keyframes. Moreover, the resulting set may include many similar images because in a TV broadcaster archive an important proportion of the assets are generated by a set of fixed cameras (studio, sports events, soap opera...), a configuration that produces several nearly duplicate keyframes. As a result, choosing all keyframes from the retrieved assets would probably result into an inefficient use of the GUI, that will be populated with many repetitive and irrelevant keyframes.

On the other extreme, selecting one keyframe per video asset may imply several problems as well. Firstly, it is necessary to correctly select this representative keyframe, a challenging task as no manual annotation is available at the keyframe scale. Moreover, the assumption that all the relevance of the video asset can be fully expressed with a single keyframe may prove wrong in many cases. Consider for

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMR '11, April 17-20, Trento, Italy

Copyright ©2011 ACM 978-1-4503-0336-1/11/04 ...\$10.00.

example a soccer match where several goals are scored. All these moments are relevant for that match and, in fact, they are normally included in every game highlights and posterior analysis. The most realistic approach is to consider that multiple keyframes may be relevant in an asset and their detection will require exploiting other techniques that will complement the manually generated annotation.

Choosing relevant keyframes is not the only design criterion to be taken into account. In the TV broadcast domain, the typical user that accesses the archives is usually interested in retrieving different video assets. The retrieved material may be used, for example, to produce new assets from previously broadcasted content or to learn about the previous treatment of a story or topic in the TV station. In general, archive users will value more the variety of video assets than retrieving several relevant keyframes from the same asset. In this context, the diversity of assets is also a requirement. Once a video asset is found, its keyframes can always be further explored in detail by applying some temporal or relevance criteria.

This paper addresses the problem of building a relevant ranked list of keyframes that belong from a diversity of assets from the archive of Catalan Broadcasting Corporation (CCMA)¹, the public TV broadcaster in Catalonia. The technique considers the visual domain as the additional modality that allows an estimation of the keyframes relevance. The proposed solution uses the visual features of the keyframes because in many cases relevant images also present similar visual features. While manual annotation is costly, these additional visual descriptors are easy to obtain because they can be automatically extracted and processed with no need of human interaction. The organization of keyframes in assets will be considered to guarantee a diversity of assets among the top hits of the results. In order to evaluate this later requirement, this paper introduces a new metric called *Average Diversity*, which is measured in parallel to the popular *Average Precision*.

The remain of this paper is structured as follows. Section 2 reviews some of the previous works that have inspired the proposed algorithm. The proposed technique is described in Section 3, while Section 4 provides experimental results on a set of data extracted from a TV archive. Finally, section 5 draws the conclusions and future work.

2. RELATED WORK

Ranking techniques offer solutions for the automatic sorting of an initial set of results according to some auxiliary criterion different from the one used at query time.

A first family of techniques are inspired in the relevance feedback techniques. Relevance feedback systems require the user interaction to determine which of the initially retrieved results are relevant for the query. In order to avoid the user presence and to obtain an automatic solution, the pseudo-relevance feedback techniques [12] consider the first results in the ranked list as relevant and the latter as non-relevant. By simulating the user interaction, the original ranked list is modified according to a relevance feedback technique and the reranked list is obtained. This type of solutions requires that the initial query generates a ranked list, a feature that may not be available in basic text-based search engines.

A second strategy for ranking is based on assessing the similarity between the elements in the initial set of results. This similarity measure can be defined on a type of feature (textual, visual, audio, social, multimodal...) and offers a wide range of possibilities for their combination. These measures are typically represented under the form of a *Similarity Graph (SG)* [3], where each node corresponds to a document in the database and each edge is weighted according to the similarity between the two connecting nodes. Once the SG is built, an estimation of the relevance of every node is obtained by considering that those nodes with more connections to other highly connected nodes are the most relevant for the query. This approach is inspired by the web search engines that sort their results according to the links that connect the web documents, such as the PageRank [7]. In the web search case, those sites referred by other important sites are considered important and are ranked first after a text query.

Once the graph-based representation of the items is generated, the structure is exploited to estimate the relevance score for every node. A popular approach to solve this problem is the *random walk*, an algorithm that identifies the document relevance with the probability of finding a traveller at the node if that traveller randomly jumps from node to node with a probability of taking a path proportional to the edge weight. Despite there is an exact solution for the algorithm, its computation requires the inversion of a very large matrix, a complex problem that is normally avoided by applying an iterative estimation according to the Power Method [4]. If available, the scores in the initial ranked lists can be naturally combined with those obtained through the random walk process by adjusting a leverage factor α . In the web search domain, the random walk is executed offline over the whole set of indexed documents so that, at query time, the retrieved ranked list can be quickly build according to these precomputed query-independent scores.

Previous works have applied the presented or similar principles for image or video retrieval. The work by Jing and Baluja [3] applied PageRank to rerank the image results obtained after a text query on the Google image search engine. Their conclusions report an increase on user satisfaction and a decrease on the amount of irrelevant results among the top hits. Hsu et al. [2] applied the random walk solution in a news broadcast archive that expanded the results obtained by a textual query to new assets connected through the SG ("context graph" in their paper). The edges on this graph were weighted by a multimodal similarity measure computed as a linear combination of textual (ASR and transcripts) and visual (salient points) similarities. In their work they studied two options to generate the SGs: consider the whole database (Full Ranking) or only those documents retrieved by the initial query (Partial Ranking). Their experimental results clearly show that the Partial Ranking solution is a much better option because the relevance of every node is clearly query-dependent. Moreover, their research also studied two connectivity options when building the SG: full connectivity and a reduced connectivity limited to the K-nearest neighbours that reduces the computation effort when solving the random walk. The reported results conclude that connectivity reduction is advisable for efficiency but if K is too small it may have a negative impact in precision performance. Another multimodal approach proposed by Richter et al. [9] combined textual and visual features to build a

¹<http://www.ccma.cat>

SG which was later filtered to reduced the impact of similar images coming from the same contributor in the context of community databases (eg. Flickr). A more formal and generic approach was proposed by Wang and Zhu [10], whose technique simultaneously balances the relevance and diversity in the retrieved ranked list. Finally, Yao et al. [13] used a different SG for every modality (visual and textual) whose values are iteratively propagated to the other SG to initialize a new random walk. This way, the two modalities are combined through a mutual and iterative exchange of information. Cao et al [1] proposed an algorithm that simultaneously reranks and clusters in two classes a collection of images. A single SG holds two random walks that compete for each node by iteratively estimating the probability scores and then assigning each node to the random walk label with whom it obtained a higher score.

The related work offers several hints about how to develop a ranking solution for the broadcast domain but, up to the author’s knowledge, none of them has combined them in a suitable way. The random walk has been reported as a valid method to estimate relevance, but several doubts arise about how to build the SG. In addition, the related cases have used SGs to fuse modalities, but non of them has worked with several SGs in the same modality to combine, for example, different types of visual descriptors. Finally, the filtering of edges as an antidote to avoid a bias to a specific user in social network can be adapted to any content database organised in sets where a diversity of these sets is desirable among the top ranked results.

3. PROPOSED SOLUTION

The presented study focuses on a broadcaster archive whose contents are stored under the form of video assets. Every asset consists of a video file, textual metadata and a set of automatically extracted keyframes. The search engine considered for this work is based on textual metadata that describes the complete asset. The broadcaster’s retrieval system presents the search results on a graphical user interface based on keyframes. As previously presented in Section 1, the goal of this work is to generate a ranked list of the keyframes associated to the video assets that have been retrieved after a textual query. This ranking must be built accord to the relevance of the keyframes as well as providing a diversity of video assets among the top hits.

Figure 1 shows the architecture elements of the system in a case where the initial text query *President Montilla* has retrieved two video assets, one from an interview (asset A) and a second one from the news program reporting about the interview (asset B). The keyframes associated to the assets are processed to create a SG for each type of available visual descriptor, for example, color and texture histograms. At the next stage, each of these SGs may be filtered to reduce some undesirable effects produced by the repetition and unbalanced amount of similar keyframes among the considered assets. The random walk algorithm is applied on every resulting SG to obtain a probability score for every image in every considered visual descriptor. Finally, these scores are fused to generate the final ranked list.

Our work introduces two contributions in the state of the art. Firstly, the fusion of diverse visual similarities on the random walk scores instead of directly on the visual metrics, a way of avoiding any scaling problem between types of visual descriptors. Secondly, we explore the filtering strategies

of some edges in the SGs to boost the diversity of assets in balance with the relevance of the top ranked keyframes.

3.1 Similarity Graphs Computation

In this paper, the nodes of the SGs represent a keyframe in the archive and the edges are weighted according to the visual similarity between two nodes. The computation of a SG poses three basic questions in our study case: (i) when to compute the distances between the keyframes, (ii) what degree of connectivity is required and, (iii) how to deal with multiple measures of visual similarity.

The calculation of visual distances is typically a computation-intensive effort that most systems perform offline. The process requires to evaluate the similarity between every pair of keyframes that are to be considered in the SG. Notice, though, that the exact topology of the SG is not known until query time, so the only SG that can be computed offline is a SG considering all keyframes in the database. This query-independent SG is read at query time to quickly build the SG that only considers those nodes contained among the initial query results. The process can be understood as a pruning of the full query-independent SG to build a query-dependent SG that keeps only the nodes and vertices retrieved in the initial query

The next question that arises is the degree of connectivity between nodes. Considering a full connectivity would result into a majority of very low weighted edges that would require an important computation effort during the random walk. In general, every keyframe in the database is considered similar to only a very small subset of other keyframes, so the full connectivity option is not necessary. Another option is setting a predefined amount of connections for every node as in the k-NN case of [2]. This approach would involve considering that all keyframes have the same amount of similar neighbours in the SG, an assumption which is wrong as visual similarity is content-depending. Besides, using a fix connectivity to all nodes would also imply serious architectural problems when reducing the query-independent SG to generate the query-dependent one, as during the pruning process the connectivity of each node will vary depending on the query. In order to solve this issue, the connectivity of the nodes is defined by setting a threshold on the similarity measure, that is, only establishing links between those nodes whose score is over a certain predefined threshold.

The remaining issue to be solved is how to combine different visual metrics defined for different types of features. Many retrieval systems use different types of features from the same or different modalities. In our case, four different visual descriptors from the MPEG-7 standard [6] were considered: Color Structure, Dominant Color, Color Layout and Texture Edge Histogram, which basically describe the color and texture distributions in every keyframe, as well as their spatial location. Each of these descriptors is defined by the MPEG-7 standard together with an individual recommendation on how to measure their similarity. Although all considered features are visual, the proposed scheme can also be applied to combine features from different modalities. Each of the four visual descriptors has an associated metric that provides a similarity criterion for ranking. In order to obtain a single and final ranked list it is necessary to combine the individual distances according to a fusion function. One option is to compute the final similarity metric by fusing the scores obtained for each feature. This solution

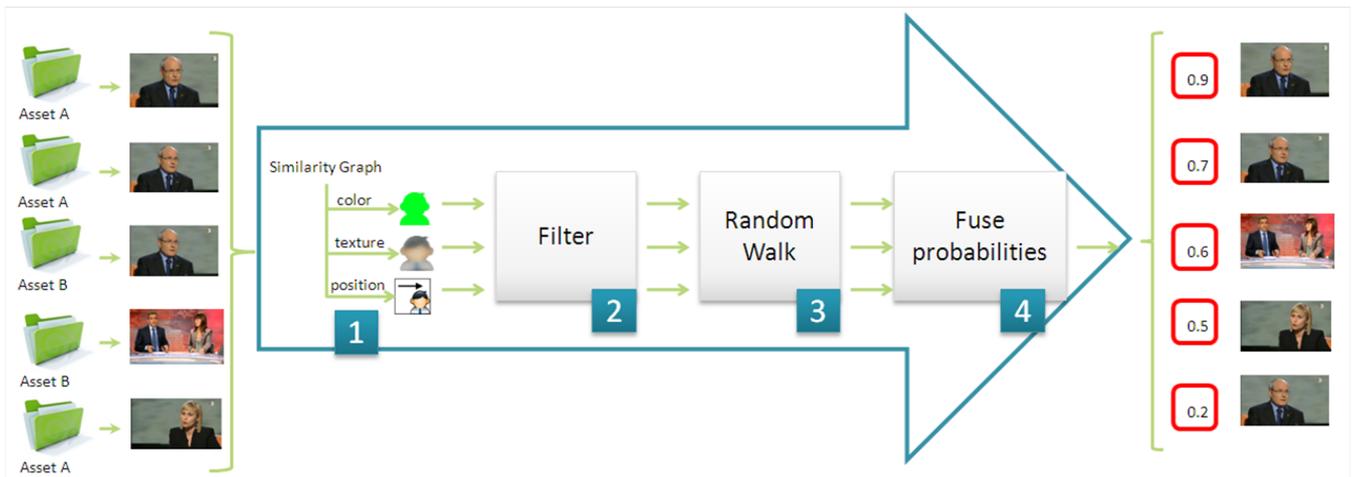


Figure 1: System architecture.

presents an important problem of combining values coming from different metrics, whose interpretation may be feature-dependent or, in other words, we may be mixing apples with pears. Another option is to delay the fusion of feature scores after the random walk, a process that generates probability scores for every node. This strategy avoids choosing any normalization strategy for the similarity metrics to safely combine probabilities. For this reason, the second option was chosen and the obtained values are fused after the random walk through an averaging operation. The fusion by averaging was chosen for simplicity because the optimization of feature weights is out of the scope of this paper. For further details on the topic, the reader is referred to [11].

3.2 Edge Filtering

Given a SG, the random walk algorithm assigns a score to each of its nodes to estimate the relevance of each keyframe. The algorithm produces higher scores for highly connected nodes, especially if their neighbouring nodes are also highly connected. Applying this approach in the context of the described broadcaster archive may produce some undesirable results that can be corrected with a previous filtering of some of the edges. This section describes which problems may arise and proposes two filtering strategies to minimize them.

The goal of the designed ranking algorithm is to boost the relevance of the keyframes and the diversity of the video assets among the top hits in the results. The main assumption is that, given a set of video assets retrieved after the text query, relevant shots will be repeated in multiple different assets. The assumption of *repetition* must be decreased to *nearly duplicate* due to the different behaviour that even the same keyframe extractor may present when dealing with edited versions of the same content appearing in different video assets.

If no post-processing is applied to a SG, the desired requirements may not be achieved due to two basic problems. Firstly, whenever an asset contains a proportional large amount of similar keyframes, these keyframes will tend to be highly scored due to their intra-connectivity, even if they do not appear in any other of the retrieved assets. For example, Figure 2 shows the expected results in the case of

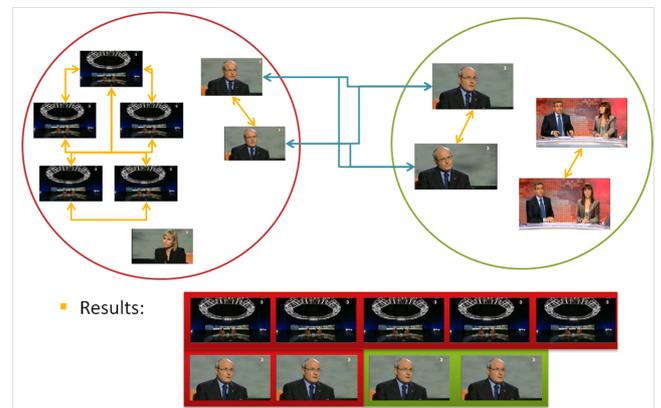


Figure 2: Irrelevant keyframes among the top hits.

the two video assets introduced in Figure 1. In this example, the keyframes showing the TV set may occupy the top hits in the ranked list despite not being the most relevant for the query. Notice that these keyframes were not even included in the news program, a proof of its irrelevance. The repetition of nearly-duplicate keyframes for the same asset is a common situation because the keyframe extractors are normally designed to generate keyframes to summarize the asset contents along the temporal dimension.

Even if the most repeated keyframes in the asset are the most relevant ones, a second problem may occur in terms of diversity. If there are many similar relevant keyframes in every asset, the top hits will be composed of blocks of relevant keyframes grouped by video assets. This situation will harm the diversity of video assets among the first positions of the list and, for this reason, should be also treated. This scenario is reflected in Figure 3, where the three first top hits include nearly-duplicate keyframes from the same asset and the second relevant asset does not appear until the fourth hit when, ideally, this should be the second one.

In order to reduce the impact of these two situations, two strategies have been defined: *intra-* and *inter-*asset filtering of the SGs. These filtering operations aim at increasing the asset diversity by preserving at the same time keyframe

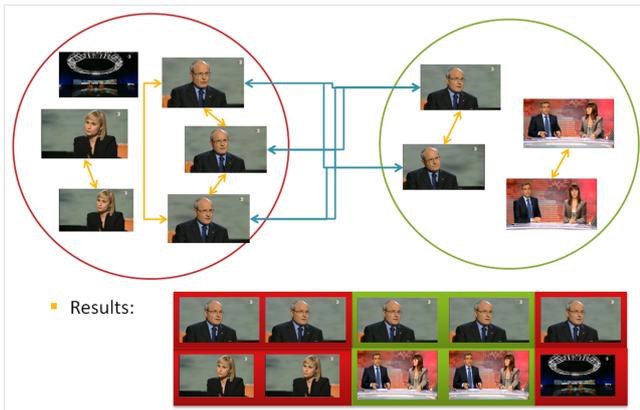


Figure 3: Blocks of assets among the top hits.

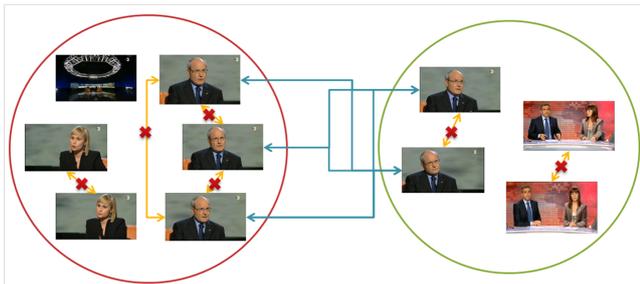


Figure 4: Intra-asset filtering.

relevance estimated by the random walk.

3.2.1 Intra-asset filtering

The main assumption of the presented reranking algorithm is that a keyframe is relevant when itself or one of its nearly-duplicates appears in different assets. This concept is applied by decreasing the scores of those keyframes whose relevance is obtained from the same asset. In practice, this reduction is achieved by deleting the edges between nodes in the same asset. This way, the only connections that a node can receive will come from an external asset, a topology that satisfies the proposed relevance definition. This operation is shown in Figure 4.

3.2.2 Inter-asset filtering

Applying an intra-asset filtering may not be enough in some cases. Even by removing the intra-asset edges in the case shown in Figure 3, several keyframes from both assets would still occupy the first positions of the ranked lists in blocks, as their relevance would be increased by multiple edges originated in the same external asset. In other words, a near-duplicate in a second asset is a good sign of relevance, but several near-duplicates in this second asset should not increase the relevance. The next degree of contribution should come from a near-duplicate from a third asset. In order to avoid excessive relevance boost from a single external asset, the amount of edges connecting every node to nodes related to another asset is limited to one. Other options could also be applied, like the normalization of the edge weights [9].

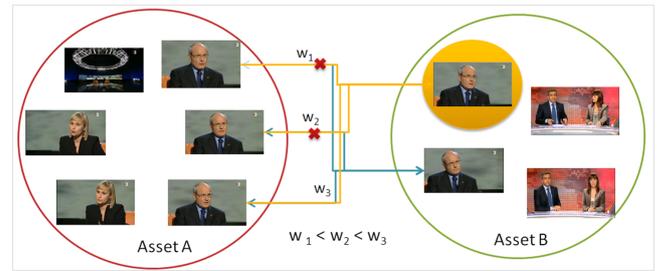


Figure 5: Inter-asset filtering.

Table 1: Test dataset

Query	# assets	# KFs
Table tennis	3	1,116
Formula 1	6	3,441
Parliament	12	2,816
Accident	8	66
Football	16	416

4. EVALUATION

The presented techniques for filtering the SGs at the intra- and inter-asset level were tested on a representative dataset from CCMA, the public national broadcaster in Catalonia. The obtained ranked lists were evaluated in terms of keyframe relevance and asset diversity for different query topics and the generated results compared.

4.1 Experimental set up

The reported experiments were based on real data corpus extracted from the broadcaster archive. The first step was the selection of a set of text queries from a list of controlled terms used by the documentalists that manually annotate the video assets. The criteria for selecting the text queries were i) being part of the annotation thesaurus used by the documentalists, ii) having multiple relevant assets in the test dataset, and iii) presenting distinctive visual features at the global image scale. For every query, all retrieved keyframes were individually annotated as relevant or not-relevant for the query. The criterion to establish the relevance was to consider if the keyframe may be tagged with the textual query by an annotator who would only access the keyframe, that is, with no knowledge about the rest of the data in the asset.

This annotation provided a ground truth over which the relevance of the retrieved keyframes was evaluated. Table 1 describes the annotation, composed of five text queries that correspond to generic concepts expressed with the controlled thesaurus used by the broadcaster documentalists. The table includes the amount of different assets and the total amount of keyframes contained in the video assets annotated with the concept keyword.

On the other hand, the whole set of keyframes retrieved by the textual query was reranked using four different MPEG-7 visual descriptors: Color Structure, Dominant Color, Color Layout and Texture Edge Histogram [6]. The random walk algorithm was initialized with uniform score for all nodes. Results were generated by comparing the generated ranked list for every filtering option and the ground truth contained in the manual annotation.

4.2 Metrics

This paper pursues the generation of results that accomplish two basic properties: relevant keyframes and diversity of assets. Two different metrics were used to evaluate these two qualities: the average precision and the average asset recall.

The *Average Precision (AP)* is broadly used by the retrieval community when evaluating the relevance of the retrieved results. This measure is obtained by averaging the first m precision values that can be obtained from the resulting ranked list as

$$\text{Average Precision}(AP) \equiv \frac{1}{m} \sum_{k=1}^m \text{Precision}(k) \quad (1)$$

where the $\text{Precision}(k)$ is the proportion of relevant keyframes when considering the first k positions in the ranked list.

The diversity of video assets has been measured with a new proposed metric inspired by the *S-recall* proposed in [14]. The S-recall stands for "subtopic recall" and measures the percentage of subtopics covered by the set of first K retrieved documents. In our study, the goal was to design a metric that behaved similarly to the AP, that is, a normalized value whose best output were the unit and that would introduce a larger penalization to the non-diverse results when they occur among the earliest positions than when they occur in the latest ones. In a first approach, the *Diversity at k* would measure the variety of the results as

$$\text{Diversity at } k \equiv D(k) = \frac{d(k) - 1}{k - 1} \quad (2)$$

where $d(k)$ corresponds to the amount of different video assets contained in the positions $1 \dots k$ of the ranked list. Notice that this metric is only defined for $k \geq 2$ as the diversity can only be evaluated on a set of multiple items.

Combining the concept of Equation 1 with the diversity measure introduced in Equation 2, we propose the *Average Diversity (AD)* as the second metric to evaluate any system where the diversity is among its specifications. The expression in 3 combines the Diversity at the k first positions, starting on 2 and going on until m , where m represents the total amount of different assets that are relevant to the query.

$$\text{Average Diversity}(AD) \equiv \frac{1}{m-1} \sum_{k=2}^m D(k) \quad (3)$$

The AD satisfies the quality of measuring 0 for homogeneous results, 1 for perfect diversity and, if a certain uniformity appears among the results, produces lower values when this uniformity is located among the first positions of the ranked list.

Both AP and AD were calculated for every text query and their values averaged among all topics to obtain the *Mean Average Precision (MAP)* and *Mean Average Diversity (MAD)*.

4.3 Results

The previous solutions for the filtering of the SGs have been evaluated by considering the four presented options for ranking: non filter (only random walk), intra-asset filtering,

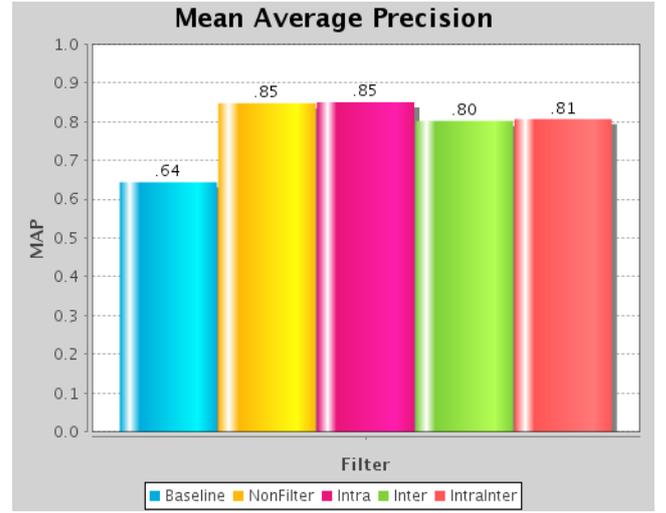


Figure 6: Mean Average Precision (MAP).

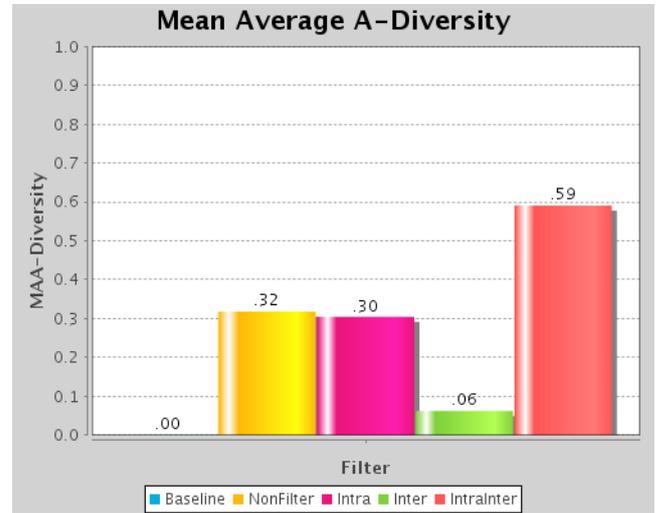


Figure 7: Mean Average Diversity (MAD).

inter-asset filtering and both types of filtering. An additional baseline case was considered by using the results list obtained after the text search, with no further processing.

The Mean Average Precision and Diversity represented in Figures 6 and 7 clearly show the relevance increase introduced by the random walk. The filtering of the SG has little impact in the MAP, with a slight decrease when the inter filtering is introduced. The decrease is reasonable as any filtering operation is an action against the principles of relevance estimation in the SG: the more relevant are the more connected and, by removing connections, there is a loss in the data used to estimate relevance. In compensation, Figure 7 proves that the filtering strategies increase the diversity of assets in the results. The removal of only inter-asset connections significantly decreases the MAP as it isolates groups of relevant keyframes whose score decreases in favour of other keyframes from their same asset. Nevertheless, the best results are obtained when the inter-asset filtering is combined with the intra-asset solution.

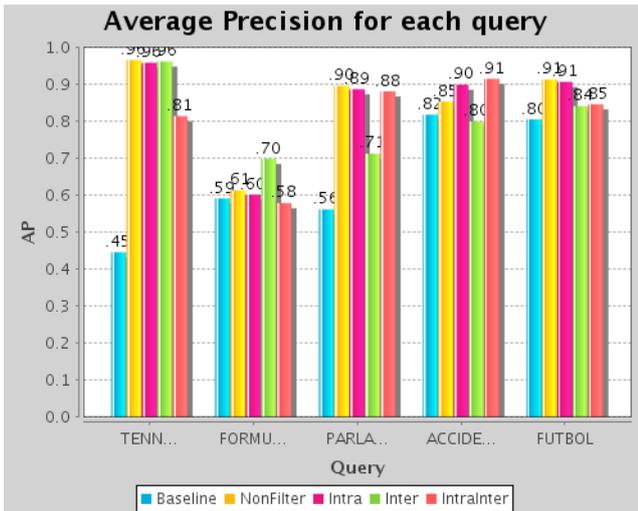


Figure 8: Average Precision for each query concept.

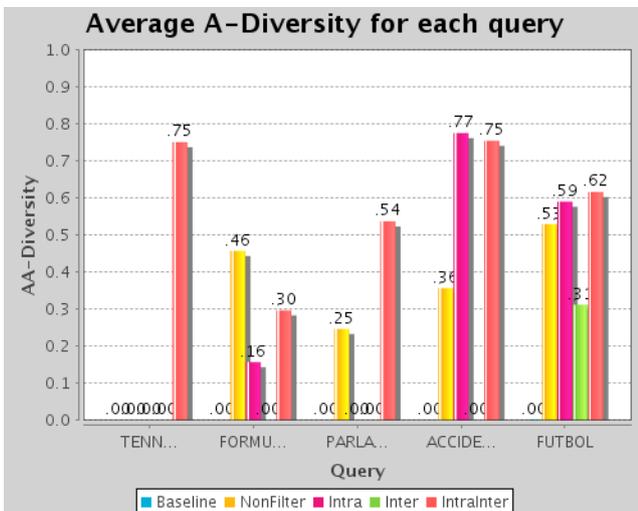


Figure 9: Average Diversity for each query concept.

The results per query concept are presented in Figures 8 and 9. The first conclusion from these figures is that the domain of application of the filtering techniques has an impact on the obtained results. While the general conclusion drawn from the MAP and MAD analysis apply, not all query concepts present the exact same behaviour. For example, the intra+inter filtering does not present the best AD in the "Formula 1" and "Accident" domains, although in general its behaviour is the most regular in terms of diversity.

The filtering stages require a computation effort to delete the edges in the SG but, on the other hand, they also simplify the iterative processing of the random walk. The experimental measures shown in Table 2 point out that there is no generic conclusion about what the final impact of the filtering is in terms of computation.

While in the case of *Accident* (few KFs) the introduction of the inter-filtering reduces the computation time compared to only intra-filtering, in the rest of the cases (more KFs) the combined inter- and intra-filtering presents faster times. As

Table 2: Computation time (in ms)

Query	NonFilter	Intra	Inter	InterIntra
T. Tennis	80,943	95,649	97,154	109,971
Formula 1	495,578	819,981	610,358	855,534
Parliament	575,740	846,400	724,752	977,864
Accident	1,277	806	734	737
Football	12,382	6,148	5,284	6,431

the computation effort for filtering will always be larger in the intra-case than in the inter & intra case, the difference is obtained at the random walk. In the *Accident* experiment, the random walk for the intra case takes longer than the sum of times of inter-filtering and the random walk over the inter- & intra-filtered SG. Although this situation has only occurred in the smallest of the experiments, not only the amount of KFs affects the random walk iteration, but also the topology of the SG.

It has also been observed that, in case of combining both intra- and inter-filtering it is advisable to perform the intra-case first because this step normally deletes more edges than the inter-case. When applied first, the intra-case reduces the amount of edges that must be checked for inter- and reduces the computation effort. Also for this reason, the obtained times for the inter & intra case do not correspond to the sum of times when applied separately.

In any case, the obtained times are too high in terms of usability. This is mainly due to the current underlying architecture, which stores the SG nodes in XML files. These files were read from a network disk and parsed at the beginning of every experiment.

5. CONCLUSIONS

This paper has presented a system architecture to implement a visual ranking solution based on the computation of the random walk algorithm over a SG. The query-dependent similarity graph has been filtered following to different strategies, intra- and inter-asset, in order to solve diversity problems of the basic technique when applied on a broadcaster archive.

The evaluation has proved the validity of the random walk approach to detect the relevance of the keyframes when the manual annotation is at the highest scale of the video assets. The repetition of certain shots in multiple assets of the archive can also be interpreted as an implicit annotation of the content in terms of relevance. This fact is exploited by the random walk algorithm to determine which keyframes are the most representative for the query.

The presented results prove that the filtering of the similarity graph is a valid technique to improve the asset diversity of the keyframe-based results. The experimentation suggests that the intra-asset filtering increases the diversity by itself, but that the inter-asset filtering only has a significant impact when combined with the intra-asset filtering. The filtering steps significantly increase the asset diversity with little impact on the precision gained during the random walk. The observations also show that the concept type and collection of assets have an impact on the overall performance, with great disparity if using intra or inter filtering separately. Nevertheless, when both filtering strategies are combined, results are more stable and generally better.

From the computation point of view, the introduction of

the filtering stages may increase the required effort as well as decrease it by building a simpler SG, this is a query-dependent behaviour. Results suggest that the computation performance may decrease for large datasets but improve in small ones.

The ranking technique has been proven as a valid solution in the domain of the considered TV broadcaster. Nevertheless, best results are to be obtained when this technique is combined with other strategies such as keyframe clustering to increase diversity, relevance feedback to learn the visual descriptors weights, user preferences or repository history.

Diversity ranking can be understood as a method for exploiting the implicit content annotation. Firstly, it uses the repetition of content in the database to detect which keyframes are the most relevant. This principle is based on the assumption that relevant material will be more often reused by the video editors generating new material for the repository. Secondly, the organization of keyframes in assets is a second type of keyframes organization that is naturally generated by at ingest time. The exploitation of implicitly generated data is a promising research line because it allows the improvement of the retrieval results without any further extra effort from the documentalists annotating the content.

The current implementation of the system is not ready yet for exploitation because of a lack of an efficient indexing strategy. The present efforts are focused in two directions: the adaptation of the Hierarchical Cellular Tree [5] into a multimodal context and an inverted index strategy based on a previous quantization of the feature space [8].

Future work includes testing the technique in the case of visual queries, which will provide an initial score that will condition the random walk scores. Another promising direction is to explore the co-ranking by mutual re-enforcement between the different SGs built for every visual descriptor.

6. ACKNOWLEDGEMENTS

All images used in this paper belong to TVC, Televisió de Catalunya, and are copyright protected. They have been provided by TVC with the only goal of research under the framework of the BuscaMedia project.

This work was partially founded by the Catalan Broadcasting Corporation (CCMA) through the Spanish project CENIT-2009-1026 BuscaMedia: "Towards a Semantic Adaptation of Multinetwork and Multiterminal Digital Media", and by the project of the Spanish Government TEC2010-18094 MuViPro: "Multicamera Video Processing using Scene Information: Applications to Sports Events, Visual Interaction and 3DTV".

7. REFERENCES

- [1] L. Cao, A. Del Pozo, X. Jin, J. Luo, J. Han, and T. S. Huang. Rankcompete: simultaneous ranking and clustering of web photos. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1071–1072, New York, NY, USA, 2010. ACM.
- [2] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 971–980, New York, NY, USA, 2007. ACM.
- [3] Y. Jing and S. Baluja. Pagerank for product image search. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 307–316, New York, NY, USA, 2008. ACM.
- [4] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 261–270, New York, NY, USA, 2003. ACM.
- [5] S. Kiranyaz and M. Gabbouj. Hierarchical cellular tree: An efficient indexing scheme for content-based retrieval on multimedia databases. *Multimedia, IEEE Transactions on*, 9(1):102–119, 2007.
- [6] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7, Multimedia Content Description Interface*. John Wiley and Sons, Ltd., Jun 2002.
- [7] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the webs. *Stanford Digital Library Technologies Project*, November 1998.
- [8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.
- [9] F. Richter, S. Romberg, E. Hörster, and R. Lienhart. Multimodal ranking for image search on community databases. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 63–72, New York, NY, USA, 2010. ACM.
- [10] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 115–122, New York, NY, USA, 2009. ACM.
- [11] L. Xie, A. Natsev, and J. Tesic. Dynamic multimodal fusion in video search. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1499–1502, 2007.
- [12] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In E. Bakker, M. Lew, T. Huang, N. Sebe, and X. Zhou, editors, *Image and Video Retrieval*, volume 2728 of *Lecture Notes in Computer Science*, pages 649–654. Springer Berlin / Heidelberg, 2003.
- [13] T. Yao, T. Mei, and C.-W. Ngo. Co-reranking by mutual reinforcement for image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 34–41, New York, NY, USA, 2010. ACM.
- [14] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '03, pages 10–17, New York, NY, USA, 2003. ACM.