Automatic Keyframe Selection based on Mutual Reinforcement Algorithm

Carles Ventura*, Xavier Giro-i-Nieto*, Veronica Vilaplana*, Daniel Giribet† and Eusebio Carasusan†
*Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
†Corporació Catalana de Mitjans Audiovisuals (CCMA), Barcelona, Spain

***Abstract*—This paper addresses the problem of video summarization through an automatic selection of a single representative keyframe. The proposed solution is based on the mutual reinforcement paradigm, where a keyframe is selected thanks to its highest and most frequent similarity to the rest of considered frames. Two variations of the algorithm are explored: a first one where only frames within the same video are used (*intra-clip* mode) and a second one where the decision also depends on the previously selected keyframes of related videos (*inter-clip* mode). These two algorithms were evaluated by a set of professional documentalists from a broadcaster's archive, and results concluded that the proposed techniques outperform the semi-manual solution adopted so far in the company.**

I. Introduction

Dealing with audiovisual and multimedia in a large scale in a business or broadcast environment requires very fast interfaces and interaction. Whenever operating vast amounts of multimedia content, any efficiency gain leads to a significant positive impact in productivity. The large volume of involved data prevents content producers, metadata annotation specialists and systems operators from watching through all the content to perform their daily activity. Instead, these users skim through vast amounts of content to select, edit and perform their professional tasks. Therefore, any visual summarization that automatic systems can provide are key to real efficiency gains and perceived user comfort.

A classic scenario for automatic content analysis is a video database. A very basic search engine would present the search results as ranked lists of textual metadata (title, author, date, etc.), which is in direct contradiction with the visual nature of the content being searched. Such listings, though practical and direct, become more usable if a still frame of each video is shown next to each result item. However, the general problem of which video frame should be presented to the user is then encountered. If randomly chosen, it may not be representative of the content. Representativeness can be guaranteed if manually chosen by editors or annotation specialists, but the manual selection requires valuable user time and its performance can degrade due to tiredness. Therefore, any system that can automatically select a relevant and distinct still frame from a video of any given length will be both time-saving and robust to human fatigue.

This paper is structured as follows. First, Section II gives an overview of the existing video summarization techniques and focuses in techniques based on the mutual reinforcement algorithm. Then, in Section III we present our approach, which is based on mutual reinforcement and consists of two different modalities: the $intra-clip$ mode and the $inter-clip$ mode. In Section IV we carry out some experiments to show that the semi-manual solution adopted so far in the broadcasting company can be replaced by our automatic proposal. Finally, in Section V we draw the conclusions.

II. Related Work

A. Video summarization techniques

Video abstraction is an integral part of many video applications, including video indexing, browsing, and retrieval. *Truong et al* [1] carried out a wide survey and review of the research in two dominant forms of the video abstraction: keyframe sets and video skims. Keyframe sets consist of a collection of salient images extracted from the underlying video source. Video skims consist of a collection of video segments (and corresponding audio) extracted from the original video.

For the keyframe sets, there are some approaches based on clustering techniques, which treat video frames as points in the feature space and work on the assumption that the representative points of clusters formed in this space can be used as keyframes for the entire video sequence. *Furini et al* proposed STIMO [2], a summarization technique designed to produce still and moving storyboards. *Lux et al* [3] and *Hadi et al* [4] also proposed a video summarization algorithm based on the k-medoid clustering algorithm to find the best representative frames.

Ciocca et al [5] proposed an algorithm which escapes the complexity of methods based on clustering and determines the complexity of the sequence in terms of changes in visual content expressed by analyzing the difference between two consecutive frames. Other approaches [6] [7] [8] are based on perceived motion energy or visual attention models.

In addition to these presented approaches, there exist other summarization techniques based on classic algorithms for graph analysis, such the random walk and the mutual reinforcement algorithms. Since our proposal is based on this approach, we give a more detailed review of these techniques in the following section.

B. Mutual reinforcement based keyframe selection techniques

The mutual reinforcement principle has been previously used for image ranking and construction of keyframe sets. *Joshi et al* [9] proposed an approach that exploits a graph structure to determine which are the most connected nodes and, by doing so, estimate their relevance in the whole structure. In these graph techniques, the entities are represented as nodes

978-1-4799-0956-8/13/$31.00 ©2013 IEEE

and the edges represent endorsements that entities give each other. They had been widely reported in literature especially in the domains of evaluation of scientific journals and on Web search, where tipically these endorsements could be thought as citations and hyperlinks respectively. It was in [9] that a set of candidate images I_1, I_2,...,I_N was assumed to form a graph with the images acting as nodes and image similarities forming the weights in the edges. They defined the rank of image I_i as r_i which is the solution to the equation:

$$r_i = \sum_{j=1}^{N} s_{ij} r_j \tag{1}$$

where s_{ij} represents the similarity between images I_i and I_j.

More recently, *Liu et al* [10] proposed an approach to generate web video thumbnails which takes account not only the video content, but also the user's query. Once keyframes have been ranked applying the mutual reinforcement algorithm, a relevance model based method is used to compute the similarity between the video frames and the user's query words. Finally, the representativeness of the keyframes in the video and the frame to query keyword relationship are linearly combined as shown in Equation 2:

$$R_i = \alpha_0 r_i + \sum_{k=1}^{K} \alpha_k s_{ki} \quad s.t. \quad \alpha_0 + \sum_{k=1}^{K} \alpha_k = 1 \tag{2}$$

where R_i is the final ranking score of frame i, r_i is the reinforcement based ranking score of frame i, s_{ki} is the similarity between the keyword k and frame i, α_0 and α_k are the weighting parameters to modulate the keyframe representativeness and text-to-keyframe relation. Once ranking scores R_i are computed, the keyframe with the highest score is selected as the video thumbnail.

He et al. [11] proposed a similar approach to determine the most representative images in an image database. This system, called ImageRank, employs the *random walk* algorithm [12] to rank the images based on the visual hyperlinks among the images. Given a graph with vertices and a set of weighted edges, the ranking scores correspond to the likelihood of arriving in each of the vertices by traversing through the graph (with a random starting point), where the decision to take a particular path is defined by the weighted edges. This graph is based on a similarity matrix modeled by low-level features between images. The main difference from the mutual reinforcement approach is that the random walk uses a probability or transition matrix, which is consequence of a normalization process performed over the similarity matrix.

A random walk based approach was also used in [13] to solve a different problem: most web search engines return images solely based on the text that surround them, rather than examining their actual visual content. *Jing et at* proposed to analyze the distribution of visual similarities among the images retrieved by the search engine and presented a system called VisualRank. *Hsu et al* [14] had also tried to solve a similar problem with the random walk algorithm.

The random walk algorithm was also used in [15] for image search. *Yao et al* proposed the idea of co-reranking the images by jointly exploring the visual and textual information. This same approach is also developed by *Wang et al* in [16], where the Joint-Rerank proposal models the images as a multigraph where each image is a node with multimodal attributes (textual and visual cues) and the parallel edges between nodes measure both image intra-modal and inter-modal similarities. *Richter et al* also used the random walk algorithm in [17], where the image ranking problem is solved on a multimodal similarity graph that combines visual features and user tags. The authors focused on finding images that most people associate with the query term.

Giro-i-Nieto et al [18] employed a variation of the random walk algorithm to obtain more relevant keyframes among the top hits of the resulted ranked lists but, at the same time, keeping a diversity of video assets in a broadcaster archive. They proposed the idea of *intra*-asset and *inter*-asset filtering to give more importance to the co-occurrences of similar frames from different assets than from the same asset.

III. Our approach

As explained in Section I, our main goal is to design an end-to-end system for a broadcasting company to automatically select the representative keyframe which will be used for the previsualization of a video retrieved by a search engine. We use the mutual reinforcement algorithm to rank the images within a video clip and to select the most representative keyframe as the top ranked. Two different techniques are proposed: (i) the *intra-clip* mode, and (ii) the *inter-clip* mode. The intra-clip mode consists on selecting the representative keyframe by using only the input video clip. On the other hand, the inter-clip mode not only uses the visual content but also the associated textual metadata. These metadata allow the retrieval of related clips with a previously selected keyframe. The retrieved keyframes provide additional criteria to select the representative keyframe for the analyzed video. In both techniques, the starting point is a visual similarity graph in which the nodes represent the images resulting from a uniform downsampling process and the edge weights represent how similar two connected images are. The final selected keyframe will be one among the selected during the downsampling. In order to compute the similarity values, a visual descriptor extracted from a global scale was adopted. In particular, we compute the MPEG-7 Color Structure Descriptor since it gives the best performance in CBIR systems according to [19]. Working at global scale should not introduce a significant bias and is significantly faster with respect to a local scale approach, according to [20].

A. Intra-clip mode

The first proposed method only exploits the visual content within the processed video. This solution assumes that the representative keyframe should correspond to the class of frame with most coverage in the video clip, i.e. the frame which is the most visually similar to the rest of frames within the video clip. Our initial approach was to apply the random walk based ranking algorithm over the similarity graph to select the representative keyframe. This algorithm can be formulated in the following way:

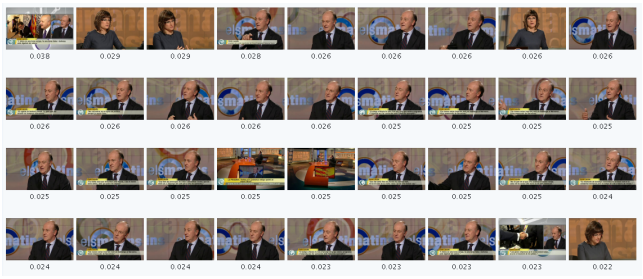

Fig. 1. Results in intra-clip mode with random walk algorithm. The values below the images represent their normalized ranking scores.

Fig. 2. Results in intra-clip mode with mutual reinforcement algorithm

$$\overrightarrow{x}(k+1) = \alpha P \overrightarrow{x}(k) + (1-\alpha)\overrightarrow{v} \tag{3}$$

where $\overrightarrow{x}(k)$ is the normalized vector which contains the ranking scores for each node at iteration k, P is the similarity or transition matrix, v is the initial ranking score vector, and α is a weighting factor. Vector $\overrightarrow{v}$ is set according to a uniform distribution since no a priori information is available. The transition matrix P is computed as the column-normalization of the similarity matrix S, which contains the similarity values between each pair of images in the frame set. However, the normalization process over matrix S can produce undesired effects as showed in Figure 1, in which the image with the highest score has least similarity with most of the video frames.

The relative decrease of the scores associated to the most frequent types of frames is due to the normalization step. During this step, the weights of the outgoing edges from every node are normalized to represent the probability for the next transition. Therefore, frames strongly connected can get their weights significantly reduced, whereas the most isolated frames may get them increased. An alternative solution to the problem is the mutual reinforcement algorithm. This algorithm is based on a similar idea, but it does not require any normalization step over the similarity matrix. Figure 2 shows the result of applying the mutual reinforcement algorithm on the same video processed in Figure 1. Now, the results match the expected ones because the type of frame with most coverage in the video clip obtains the highest score followed by other frames which are not so frequent in the video.

Therefore, for the intra-clip mode we apply the following algorithm [9]:

1) Initialize $\overrightarrow{r^0} = (r_1^0, r_2^0, ..., r_N^0)$ randomly such that $\sum_{i=1}^{N} r_i^0 = 1$ and $r_i^0 > 0 \forall i$
2) Assign 1 to t
3) $r_i^t = \sum_{j=1}^{N} s_{ij} r_j^{t-1} \ \forall i \in 1, ..., N$
4) Assign $\frac{r_i^t}{||\overrightarrow{r^t}||}$, where $||\overrightarrow{r^t}|| = \sum_{i=1}^{N} r_i^t$
5) Assign $t+1$ to t
6) Repeat steps 3 to 5 till convergence

where r_i^t is the rank score of image I_i at iteration t, s_{ij} represents the similarity between frames I_i and I_j, and N is the number of frames resulting from the uniform sampling process. In our experiments, we consider that convergence is achieved when the obtained rank list does not change or a maximum amount of 20 iterations is reached.

One of the requirements of system imposed by the broadcaster company was that the selected representative keyframe should not include any textual caption. Therefore, a post-processing step was added, which consists of a text filtering. This text filtering, proposed in [21], is based on setting a maximum threshold over the energy coefficients in the different subbands (LH, HL and HH) of the Haar-wavelet transform computed on the lower part of the image, where caption text normally appears.

B. *Inter-clip mode*

The selection of the representative keyframe can also be based on existing knowledge. Therefore, the criteria of maximum coverage is complemented by the prior knowledge provided by pre-analyzed videos from which the representative keyframes are available. Thus, in the inter-clip mode, the automatic selection process depends on the input video clip as well as the representative keyframes selected for videos stored in the database. We have considered two different methods to retrieve the related videos:

- *Textual search*: Some keywords extracted from the input video metadata are used to retrieve the videos whose metadata include the query keywords. Furthermore, since the videos are categorized, the results of a search can be filtered by category.
- *TF-IDF descriptors* [22]: This approach employs a set of training metadata files to build a vocabulary, from which the stop words were removed. Every word from the vocabulary has an associated value which represents how often that term appears with respect to the whole set of documents. This Document Frequency (DF) is inverted (IDF) and combined with the count of terms in each document in the target dataset (Term Frequency-TF). The result is a fixed-length feature vector that allows comparing text documents according to the entropy of the contained words.

After retrieving these related videos, the MPEG-7 Color Structure Descriptors are extracted from their representative keyframes. Then, for each input video frame I_i the mutual reinforcement algorithm is applied as in the intra-clip mode to obtain their scores r_i. Once the scores have been computed without considering the retrieved videos, the final ranking score R_i of the frame I_i is computed by linear fusion:

$$R_i = \alpha_0 r_i + \sum_{m=1}^{M} \alpha_m s(I_i, K_m) \tag{4}$$

where M is the number of retrieved videos, $s(I_i, K_m)$ is the similarity value between the input video frame I_i and the representative keyframe K_m from the mth retrieved video, and α_0 and α_m are the weighting parameters used to modulate the reinforcement based ranking score r_i and the similarity between the frame I_i and the set of representatives keyframes from the retrieved videos. Figure 3 shows the reinforcement based ranking score obtained in a video clip in the intra-clip mode and Figure 5 shows the final ranking scores for the same video in the inter-clip mode using the representative keyframes showed in Figure 4. Thus, in the inter-clip approach, the frames from the analyzed video which are similar to the retrieved representative keyframes got their scores increased.

The weighting parameters from Equation 4 are set in the following way:

- If the M related videos have been retrieved using the textual searcher, then we adopt a uniform distribution, i.e. $\alpha_m = (1 - \alpha_0)/M \ \forall m$.
- If the TF-IDF have been used to retrieve the M videos and the textual similarity between the video v_m and the input video v_{input} is $s_t(v_m, v_{input})$, then $\alpha_m = (1 - \alpha_0)s_t(v_m, v_{input})/N_t$, where N_t is the normalizing factor over the textual similarity values, i.e. $N_t = \sum_{m=1}^{M} s_t(v_m, v_{input})$.

As in the intra-clip mode, there is also a post-processing step to filter the frames containing textual captions.

Fig. 3. Mutual reinforcement based ranking score in the intra-clip mode.

Fig. 4. Representative keyframes from the videos retrieved in the inter-clip mode

Fig. 5. Final ranking score in the inter-clip mode.

IV. Experiments

The presented technique was tested in a broadcasting corporation with a large video database. The reported experiments aimed at deciding whether the automatic selection of a keyframe could improve an existing manual-based method.

A. Set up

The considered dataset is composed of a large collection of video assets, being a *video asset* the combination of a video file, textual metadata and a representative keyframe. The goal of the system under test is to automatically generate this representative keyframe given only the video file and the text metadata.

1) Compared approaches: The experiments were designed to compare four different strategies for the selection of a representative keyframe. These four options are the following:

- **Semi-manual**: The existing solution at the broadcaster company requires the documentalist to choose among three frames from the video, which have been extracted after an arbitrary time-based sampling.
- **Intra**: The keyframe is automatically selected following the *intra* technique presented in Section III-A.
- **Inter**: The keyframe is automatically selected following the *inter* technique presented in Section III-B.
- **Random**: The keyframe is automatically selected after a random sampling of the video file.

The goal of the study is to discern the relative performance of the four techniques in terms of user satisfaction.

2) User-based evaluation: A first approach for evaluation could consider a benchmark built from those keyframes which have been previously chosen by the documentalists. The experiment could run the proposed algorithms and compare the results with these ground truth keyframes. However, the evaluation of a video summary with a single keyframe poses two challenges which would not be taken into account with this set up. First, the multiplicity of keyframes that could be considered as representative of a video asset. Second, the high degree of user subjectivity when determining if a keyframe is good enough to summarize the contents of a video. These two drawbacks limit the usefulness of the results obtained with this approach.

An alternative solution is to adopt a user-based evaluation, as it was considered in this work. According to [1], this is probably the most useful and realistic form of evaluation, especially when keyframes are extracted for user-based interactive tasks such as content browsing and navigation. However, it has not been widely employed due to the difficulty in setting it up. In the adopted approach, the automatically selected keyframes are evaluated by users, in our case, the documentalists from the broadcaster company. They possess prior knowledge about the video asset and the topic it represents, so they are qualified professionals who can determine whether a keyframe is valid, considering multiple criteria inherent to their specialist job. This type of evaluation is also costly, as it requires expert users but, given the context of the project, these type of users were available and willing to collaborate.

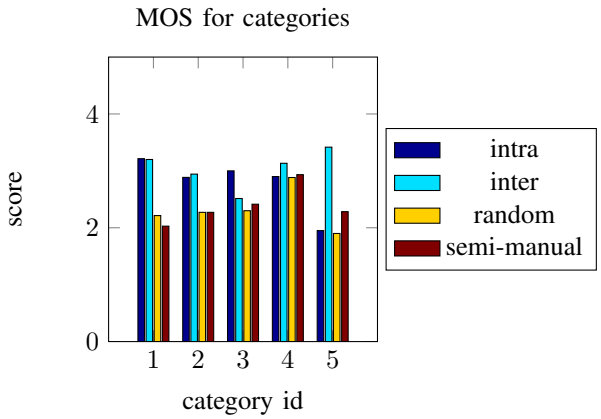

Fig. 6. MOS values for categories: (1) News-Economy, (2) News-International, (3) News-Politics, (4) Morning Show-Interview, and (5) Morning Show-Discussion

Every participating user was asked to complete 50 online polls. In each of them, the textual metadata of an asset was shown (title and description) together with four keyframes selected with each of the the four considered techniques. Users were asked to rate each of the keyframes in a 5-star scale with the following interpretation: (1) *Unacceptable*, (2) *Fair*, (3) *Good*, (4) *Very good*, (5) *Excellent*.

A diverse dataset of videos was considered to test the presented techniques in different contexts, all of them of high interest for the broadcasting company. Two main domains were selected: *Morning Show* and *News*. The first one contained images of a controlled environment, a TV studio, with a high repetition of similar shots between different assets and the clips were from 22 to 80 minutes long. On the other hand, the *News* domain was much more challenging, as its videos were much more diverse both visually and semantically, but they were shorter (from 1 to 4 minutes long). The *Morning Show* domain was split in two categories: (i) *Interview*, and (ii) *Discussion*; and the *News* domain in three categories: (i) *Economy*, (ii) *International*, and (iii) *Politics*. Each user assessed 10 assets from each category, which sums up to the 50 mentioned polls.

B. Results

A total of 7 users (documentalists) answered the 50 online polls. The evaluation of the different summarization techniques considered the Mean Opinion Score (MOS) test, which is a widely used measure of the system quality by averaging the ratings given by the evaluating users.

From a global analysis of the results, the inter-clip approach gives the best performance in terms of user satisfaction (MOS = 3.04), followed by the intra-clip strategy (MOS = 2.79). Therefore, both proposed approaches perform significantly better than the current semi-manual technique employed by the broadcasting company (MOS = 2.39) and the random strategy (MOS = 2.31). Next, the results are also analyzed for each domain and category in more detail.

The results in Figure 6 show that the inter-clip mode perfoms better than the intra-clip mode for the *Morning Show* domain, but this behavior is inverted in the *News* domain. The reason is that the videos from the *News* domain are much more diverse, and that the representative keyframes from the videos retrieved by textual similarity are not visually similar enough to the frames of the analyzed video. As a result, there is little influence in the ranking scores despite the linear fusion. Nevertheless, the evaluation of the intra-clip mode in the *News* domain is reasonably satisfactory since a score of 3.03 is obtained by averaging the MOS values for each category. Furthermore, the results of both automatic strategies are significantly better than the MOS values obtained by the random and semi-manual strategies, which have an averaged MOS of 2.26 and 2.24 respectively.

Regarding the *Morning Show* domain, each category needs to be independently analyzed.

The *Interview* category shows little variation of the visual content because of a very controlled environment. In the videos from this category, the frames showing the interviewee are the most frequent ones. Therefore, according to the maximum coverage criteria inherent in the mutual reinforcement algorithm, the intra-clip mode (MOS = 2.90) achieves the purpose of automatic selection this kind of frame. Furthermore, the inter-clip strategy (MOS = 3.13) gives more consistency to the results obtained by the intra-clip technique since the representative keyframes from the retrieved videos are visually similar, despite the fact that the interviewees differ from one video to another. However, the MOS values obtained in this category are the most similar independently of the technique applied (MOS values are 2.88 and 2.93 for random and semi-manual approaches respectively). A more detailed analysis of the users' evaluation in this domain indicates that this fact can be consequence of: (i) the user assesses the facial expression and the head pose (frontal poses are better rated), (ii) some fading captions are not detected by the wavelet-based textual filtering, and (iii) the high probability that the random approach selects a frame with the interviewee.

The *Discussion* category is a more challenging category due to the high diversity of the visual content within the same video clip. Figure 4 shows some of the representative keyframes from different video clips which look very similar. However, these types of frames are not frequent in the video content and, therefore, do not respond to the maximum coverage criteria applied by the mutual reinforcement algorithm in the intra-clip mode. This is the reason why the intra-clip and the random approaches achieve a poor MOS (1.95 and 1.90 respectively). On the other hand, thanks to the co-occurrences among representative keyframes from different videos, the inter-clip approach achieves a very good performance (MOS = 3.42), even significantly better than the semi-manual approach (MOS = 2.28).

In addition to the MOS analysis, the rate of success is also examined, i.e. how often each technique selects one representative keyframe which is acceptable for the user, independently of the specific grade from 2 (Fair) to 5 (Excellent) given to it. Figure 7 shows the acceptance rate for each technique in each category. From this figure, the same conclusions can be drawn with respect to the different approaches over the different domains and categories. The intra-clip approach achieves an averaged acceptance rate of 80.95% in the *News* domain and the inter-clip approach achieves acceptance rates of 86.67% and 93.33% in the *Interview* and the *Discussion* categories respectively.

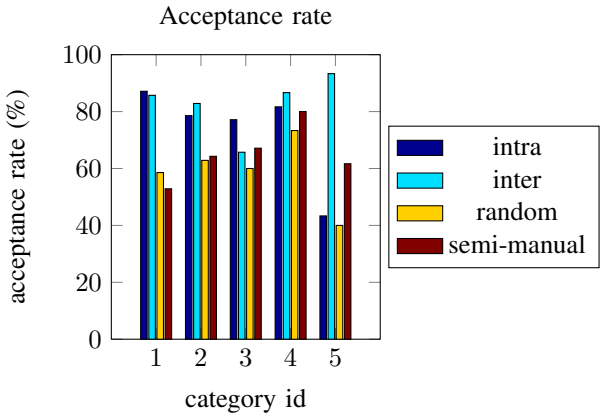

Fig. 7. Acceptance rate for categories: (1) News-Economy, (2) News-International, (3) News-Politics, (4) Morning Show-Interview, and (5) Morning Show-Discussion

V. Conclusions

In this work, we have presented a system that automatically extracts the representative keyframe from a video. The proposed system is based on the mutual reinforcement algorithm to identify the frame with maximum coverage within the video in the intra-clip approach. Furthermore, we have also proposed the inter-clip approach, in which related videos are retrieved based on their textual metadata. In the inter-clip mode, the visual similarity between the representative keyframe of the retrieved videos and the frames from the analyzed video are linearly fused with the scores provided by the intra-clip approach. The system has been evaluated in terms of user satisfaction in comparison with the semi-manual approach employed by the broadcasting company and a random approach that works as baseline. This evaluation has been carried out by the own broadcaster's documentalists, who are experts in these summarization tasks. From this evaluation, we can conclude that the current semi-manual system can be replaced by the automatic approach and that the inter-clip approach is useful and outperforms the intra-clip method in controlled environments, where the inter-clip mode can exploit the co-occurrences among keyframes from different videos.

Acknowledgment

This work was partially founded by the Spanish project CENIT-2009-1026 BuscaMedia, TEC2010-18094 MuViPro project of the Spanish Government and FPU-2010 Research Fellowship Program of the Spanish Ministry of Education.

References

[1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, Feb. 2007.

[2] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "Stimo: Still and moving video storyboard for the web scenario," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 47–69, Jan. 2010.

[3] M. Lux, K. Schöffmann, O. Marques, and L. Böszörmenyi, "A novel tool for quick video summarization using keyframe extraction techniques," in *Proceedings of 9th Workshop on Multimedia Metadata(WMM'09), CEUR Workship Proceedings*, 2009.

[4] Y. Hadi, F. Essannouni, and R. O. H. Thami, "Video summarization by k-medoid clustering," in *Proceedings of the 2006 ACM symposium on Applied computing*, ser. SAC '06. New York, NY, USA: ACM, 2006, pp. 1400–1401.

[5] C. Gianluigi and S. Raimondo, "An innovative algorithm for key frame extraction in video summarization," *Journal of Real-Time Image Processing*, vol. 1, pp. 69–88, 2006.

[6] T. Liu, H.-J. Zhang, and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 13, no. 10, pp. 1006–1013, Oct. 2003.

[7] J.-L. Lai and Y. Yi, "Key frame extraction based on visual attention model," *J. Vis. Comun. Image Represent.*, vol. 23, no. 1, pp. 114–125, Jan. 2012.

[8] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, 2012.

[9] D. Joshi, J. Z. Wang, and J. Li, "The story picturing engine: finding elite images to illustrate a story using mutual reinforcement," in *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, ser. MIR '04. New York, NY, USA: ACM, 2004, pp. 119–126.

[10] C. Liu, Q. Huang, and S. Jiang, "Query sensitive dynamic web video thumbnail generation," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, sept. 2011, pp. 2449 –2452.

[11] X. He, W.-Y. Ma, and H. Zhang, "Imagerank: spectral techniques for structural analysis of image database," in *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 2*, ser. ICME '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 25–28.

[12] L. Lovász, "Random walks on graphs: A survey," *Combinatorics, Paul erdos is eighty*, vol. 2, no. 1, pp. 1–46, 1993.

[13] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1877 –1890, nov. 2008.

[14] 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*, ser. MULTIMEDIA '07. New York, NY, USA: ACM, 2007, pp. 971–980.

[15] 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*, ser. CIVR '10. New York, NY, USA: ACM, 2010, pp. 34–41.

[16] G. Wang and X.-S. Xu, "Joint-rerank: a novel method for image search reranking," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ser. ICMR '12. New York, NY, USA: ACM, 2012, pp. 37:1–37:8.

[17] 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*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 63–72.

[18] X. Giro-i Nieto, M. Alfaro, and F. Marques, "Diversity ranking for video retrieval from a broadcaster archive," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ser. ICMR '11. New York, NY, USA: ACM, 2011, pp. 56:1–56:8.

[19] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7, Multimedia Content Description Interface*, B. S. Manjunath, P. Salembier, and T. Sikora, Eds. John Wiley and Sons, Ltd., Jun 2002.

[20] M. Kogler, M. del Fabro, M. Lux, K. Schoeffmann, and L. Boeszoermenyi, "Global vs. local feature in video summarization: Experimental results," December 2009.

[21] M. Leon, V. Vilaplana, A. Gasull, and F. Marques, "Caption text extraction for indexing purposes using a hierarchical region-based image model," in *Proceedings of the 16th IEEE international conference on Image processing*, ser. ICIP'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1849–1852.

[22] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.