Movie Shot Selection Preserving Narrative Properties

Ioannis Mademlis[†], Anastasios Tefas[†], Nikos Nikolaidis[†] and Ioannis Pitas[†⋆]

[†]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[⋆]Department of Electrical and Electronic Engineering, University of Bristol, UK

***Abstract*—Automatic shot selection is an important aspect of movie summarization that is helpful both to producers and to audiences, e.g., for market promotion or browsing purposes. However, most of the related research has focused on shot selection based on low-level video content, which disregards semantic information, or on narrative properties extracted from text, which requires the movie script to be available. In this work, semantic shot selection based on the narrative prominence of movie characters in both the visual and the audio modalities is investigated, without the need for additional data such as a script. The output is a movie summary that only contains video frames from selected movie shots. Selection is controlled by a user-provided shot retention parameter, that removes key-frames/key-segments from the skim based on actor face appearances and speech instances. This novel process (Multimodal Shot Pruning, or MSP) is algebraically modelled as a multimodal matrix Column Subset Selection Problem, which is solved using an evolutionary computing approach.**

Keywords*—*Video Summarization, Shot Selection, Column Subset Selection Problem

I. Introduction

The importance of *video summarization* has increased in recent years, due to the fact that large volumes of video data have become available. Thus, the need has arisen for automatic or semi-automatic methods for deriving short video summaries containing the most important/salient parts of the original, full-length video data. Such summaries are helpful both to video producers, by facilitating the creation of promotional material or the archival of massive video content, and to consumers, by enabling easy video browsing and search operations, e.g., in on-line video galleries [1] [2] [3]. Relevant algorithms typically attempt to balance different needs, such as representativeness, outlier inclusion, compactness (defined in this context as lack of redundancy) and conciseness. The summaries can be derived either in the form of a set of extracted key-frames (*static summarization*), or of a *video skim*, i.e., a clip made from temporally ordered concatenated key-segments, with each key-segment containing multiple sequential video frames (*dynamic summarization*).

In the special case of movies, the existence of shot cuts facilitates summarization by segmenting the video into naturally separated parts. Typically, shot boundaries are automatically extracted in a pre-processing stage and are subsequently fed into the main summarization process. Several alternative approaches may be followed for the latter, e.g., extracting one key-frame per shot by processing low-level video descriptors (and, optionally, temporally expanding it to a key-segment centered on the key-frame, for skim construction) [4] [5] [6] [7], filtering out monochrome key-frames which convey no useful information, detecting similar key-frames along the full-length video and removing them to reduce skim redundancy, pruning entire shots based on high-level characteristics or the estimated movie structure, etc. Different combinations of these approaches may also be employed.

High-level narrative properties, such as character appearance, prominence and interactions, are ideal for shot selection, since they provide a semantically meaningful structure that partly determines audience attention, according to conventional film theory [8]. Despite this ability to extract attractive movie content, narrative-preserving shot selection has not been thoroughly investigated in the summarization literature. An exception can be found in [9], where scene segmentation and a character-oriented story flow graph are exploited for summarization adhering to the narrative. However, the method requires the movie script to be available, along with the video itself.

This work presents a narrative-preserving shot selection process for movies, which does not need additional data beyond the film itself. The proposed Multimodal Shot Pruning (MSP) method discards key-segments from the derived video skim, based on which shot they belong to and on pre-existing information about temporal speech (audio) and face (visual) appearance segments. This process is regulated by a user-provided shot retention parameter and algebraically modeled as a multimodal matrix Column Subset Selection Problem (CSSP) [10], which is solved using an evolutionary computing approach. Thus, a shorter skim is produced in a systematic manner that considers the narrative prominence of movie actors.

The remainder of this paper is organized in the following way. Section II presents the proposed method, including the problem modelling approach and the evolutionary technique employed to solve it. The method can be attached to any film summarization pipeline as a shot selection step, provided that shot boundaries have been extracted. Section III describes subective evaluation experiments conducted in the context of the stereoscopic movie summarization pipeline presented in [11], in order to evaluate the performance of the proposed method. In Section IV conclusions are drawn from the preceding discussion.

II. Multimodal Shot Pruning (MSP)

Below, segmentation of the movie into shots and a film summarization pipeline that has produced a preliminary summary, are assumed given. The preliminary summary is assumed to be a temporally ordered set of key-segments (for dynamic summarization), but the exact same technique could also be

applied to the alternative case of a key-frame set (for static summarization).

The proposed Multimodal Shot Pruning (MSP) method is a post-processing shot selection step that picks shots for inclusion into the final summary, thus producing a shorter skim which still contains most of the informational content found in the preliminary one. It operates by "discarding" shots in a systematic manner that considers actor-oriented narrative properties, such as "Who spoke when?" (speakers) and "Who appeared when?" (faces), namely speaker and actor diarization information. As in the case of speakers, each face appearance consists simply of a video segment that starts and ends at the temporal boundaries of an uninterrupted face appearance. Such data may have been acquired through the successive application of face detection [12], face tracking [13], face clustering [14] and label propagation [15] algorithms. Despite these algorithmic prerequisites, no extra data modalities (such as the movie script) are required, beyond the film itself.

MSP is algebraically modelled as a matrix Column Subset Selection Problem (CSSP) [10], which is briefly discussed here. Assuming a low-rank $M \times N$ matrix $\mathbf{D}$ and a parameter $C < N$, CSSP consists in selecting a subset of exactly C columns of $\mathbf{D}$, which will form a new $M \times C$ matrix $\mathbf{C}$ that captures as much of the information contained in the original matrix as possible. The goal is to construct a matrix $\mathbf{C} \in \mathbb{R}^{M \times C}$ such that the quantity

$$\|\mathbf{D} - (\mathbf{C}\mathbf{C}^{+})\mathbf{D}\|_F \tag{1}$$

is minimized. In the above, $\|\cdot\|_F$ is the Frobenius matrix norm and $\mathbf{C}^{+}$ is the pseudoinverse of $\mathbf{C}$. Thus, the approximation of $\mathbf{D}$ by the smaller matrix $\mathbf{C}$ is expressed in terms of the Frobenius norm in a projection sense: as $\mathbf{D}$ does not have full rank, $\mathbf{C}\mathbf{C}^{+}$ is not simply an identity matrix, but acts as a projection matrix onto the span of the C columns contained in $\mathbf{C}$.

In data analysis, CSSP is an obvious choice for mathematically modelling a feature selection process as an optimization problem. It can be optimally solved by exhaustive search in $\mathcal{O}(N^C)$ time [10], which clearly is a very impractical approach. Thus, approximate algorithms with lower computational complexity have been presented in the relevant literature, with the goal of finding a suboptimal but acceptable solution. The proposed approaches include randomized, deterministic or hybrid methods, using SVD sparse approximation [16], random selection of matrix columns, based on a probability distribution, which is later refined deterministically [10], greedy recursive computation of the reconstruction error, initialized with random projections of the matrix columns [17], column subset selection with probabilities proportional to the squared volumes of the parallelepipeds defined by these subsets [18], etc.

In [19], a metaheuristic approach based on a genetic algorithm is successfully employed for the approximate solution of the CSSP, by directly using Equation (1) as a fitness function. The method is evaluated on several small, randomly generated matrices and is shown to produce good results for a fixed small value of C. In this work, the same metaheuristic approach was adopted and adapted into the proposed pipeline, so that MSP could be modelled and solved as a CSSP.

Specifically, two low-rank, sparse, binary matrices are constructed: $\mathbf{S}, \mathbf{F} \in \mathbb{R}^{V \times S}$, where S is the total number of movie shots and V is the total number of visible speakers, i.e., it is the cardinality of the intersection of the set of all visible faces and the set of all speakers. Typically, $S >> V$. Since temporal speech segment and face appearances are assumed given, $\mathbf{S}$ and $\mathbf{F}$, also referred to as *shot matrices* hereafter, are being filled with binary values:

$$\mathbf{S}_{ij} = \begin{cases} 1, & \text{if the } i\text{-th actor speaks in the } j\text{-th shot,} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{F}_{ij} = \begin{cases} 1, & \text{if the } i\text{-th actor appears in the } j\text{-th shot,} \\ 0, & \text{otherwise.} \end{cases}$$

where $1 \leq i \leq V, 1 \leq j \leq S$.

Subsequently, $\mathbf{S}$ and $\mathbf{F}$ are modified in a Gaussian expansion process, so that each speech / face appearance is "extended" to neighbouring shots. Thus, the initially binary shot matrices are converted to real ones, in a manner that preserves relevant information. That is, for each $\mathbf{S}_{ij} = 1$ / $\mathbf{F}_{ij} = 1$, a discrete approximation of a Gaussian distribution, having its peak at $\mathbf{S}_{ij}$ / $\mathbf{F}_{ij}$, is locally assigned to the entries of the i-th row around $\mathbf{S}_{ij}$ / $\mathbf{F}_{ij}$, respectively. The standard deviation of the selected probability mass function is chosen so that each appearance is extended only to the dv shots immediately preceding it and following it. Subsequently, shot matrix values derived from different speech / face appearances and corresponding to the same shot matrix entry are added, thus enabling a diffusion of neighboring speech / face appearances. dv may take any value that is less than half the average scene duration (in shots).

This Gaussian expansion process allows a rudimentary form of scene modeling, based on the provided shot segmentation and actor-oriented narrative information. It was included in order to implicitly assist the discrimination between more and less narratively prominent actors. That is, the basis vector sets of the initial shot matrices most likely coincide with the standard basis, with one basis vector corresponding to each visible speaker. However, after the Gaussian expansion, the basis vector sets of the final shot matrices most likely include basis vectors corresponding to the most prominent visible speakers and basis vectors corresponding to combinations of more and less prominent visible speakers. For instance, if the k-th visible speaker is a supporting actor that speaks / appears in scenes (and, therefore, in neigboring shots) only along with a lead actor, there will be no column vector $\mathbf{c}$ of the shot matrices where $\mathbf{c}_i = 0, i \neq k$ and $\mathbf{c}_k \neq 0$.

After the final $\mathbf{S}$ and $\mathbf{F}$ matrices have been constructed, they are implicated in a joint column subset selection problem regulated by a parameter $C = \lfloor S\frac{p}{2} \rfloor$. S is the total number of movie shots and the user-provided retention parameter p regulates the aggressiveness of shot elimination during this stage. By solving this problem, only key-segments belonging to an optimal subset of shots will be selected to appear in the final video skim, with subset optimality expressed in terms of discarding shots which correspond to shot matrix columns that are linear combinations of other columns. Thus, it is more likely to retain shots where lead actors, or combinations of

supporting and lead actors, are present, rather than supporting actors alone.

The desired solution is a set of matrix column indices with cardinality equal to C. Since $\mathbf{S}, \mathbf{F} \in \mathbb{R}^{V \times S}$, for the k-th such index with an assigned value g_k the following hold:

$$k \in \mathbb{N}, \quad k \in [1, \cdots, C]. \tag{2}$$

$$g_k \in \mathbb{N}, \quad g_k \in [1, \cdots, S]. \tag{3}$$

A genetic algorithm is employed to approximate an optimal solution and, as in [19], each candidate is encoded in the form of a sequence of column indices sorted in increasing order. Every such chromosome is of length C. Roulette selection at each iteration is adopted as the mating pool formation strategy. Assuming f_l is the evaluated fitness of the l-th candidate in the current population, this method assigns a selection probability $p^l_{sel} = f_l / \sum_{m=1}^{N} f_m$ to the l-th chromosome.

An order-preserving variant of 1-point crossover [19] is utilized as the main genetic operator. Specifically, in order to combine parent chromosomes $\mathbf{c}^l$ and $\mathbf{c}^m$, a random position k is selected as crossover point and is inspected for suitability. k is considered to be suitable as a crossover point, if the following condition holds:

$$(\mathbf{c}^l_k < \mathbf{c}^m_{k+1}) \wedge (\mathbf{c}^m_k < \mathbf{c}^l_{k+1}). \tag{4}$$

In case Equation (4) does not hold for position k, a different position is selected and inspected. This process continues until either a suitable crossover point has been detected, or all possible positions have been deemed unsuitable. In the former case, crossover is applied and the two parent chromosomes are replaced by their offspring. In the latter case, each of the implicated chromosomes is passed unaltered to the population of the next generation with probability p^l_{sel} or p^m_{sel}, respectively. If $\mathbf{c}^l$ or $\mathbf{c}^m$ is not being retained, it is replaced in the next generation by a copy of the fittest current candidate $\mathbf{c}^n$ with probability p^n_{sel}. If $\mathbf{c}^n$ is also not selected for retention, the process continues with the second fittest of the current candidates, and so on, until a chromosome has been selected.

An order-preserving variant of mutation [19] is employed as the second genetic operator. Specifically, the k-th gene of a chromosome $\mathbf{c}^n$, with an assigned valued of $\mathbf{c}^n_k$, is randomly selected and replaced by a value determined by the neighbouring genes, according to Equation (5):

$$\mathbf{c}^n_k = \begin{cases} rand(0, \mathbf{c}^n_{k+1}), & \text{if} \quad k = 1 \\ rand(\mathbf{c}^n_{k-1}, \mathbf{c}^n_{k+1}), & \text{if} \quad k \in (1, C) \\ rand(\mathbf{c}^n_{k-1}, S+1), & \text{if} \quad k = C. \end{cases} \tag{5}$$

where $rand(a, b)$ uniformly selects a random integer from the interval (a, b). Although this operator ensures a proper ordering of the indices, it has no effect when $\mathbf{c}^n_{k-1}$, $\mathbf{c}^n_k$ and $\mathbf{c}^n_{k+1}$ are successive integers.

The employed fitness function is derived from Equation (1), which is applied to both matrices $\mathbf{S}$ and $\mathbf{F}$. The matrix column indices encoded in the chromosome $\mathbf{c}^n$ which is under evaluation, give rise to the matrices $\mathbf{C}^S$ and $\mathbf{C}^F$, respectively. The former contains a subset of the columns of $\mathbf{S}$, while the latter contains a subset of the columns of $\mathbf{F}$. Thus, the fitness function that needs to be maximized can be expressed as:

$$fit(\mathbf{c}^n) = \frac{1}{\|\mathbf{S} - (\mathbf{C}^S \mathbf{C}^{S+})\mathbf{S}\|_F + \|\mathbf{F} - (\mathbf{C}^F \mathbf{C}^{F+})\mathbf{F}\|_F}. \tag{6}$$

Once the described genetic algorithm has converged to a solution $\mathbf{c}^{best}$, all key-segments belonging to shots not encoded (by their corresponding column index) in $\mathbf{c}^{best}$ are removed from the produced movie skim.

III. Evaluation

For the purpose of evaluating the proposed method, MSP was integrated as a shot selection step in the stereoscopic movie summarization pipeline presented in [11], which sequentially applies shot cut detection, key-frame extraction per shot, movie-wide similar key-frame filtering, key-frame extension to key-segments, key-segments temporal concatenation and stereoscopic visual defects elimination. MSP was applied just before the last step of that pipeline. Parameter p was employed both for the proposed MSP stage and for the movie-wide similar key-frame filtering stage of the pipeline (where it determines the number of clusters, as stated in [11]).

A subjective evaluation scheme was employed, similar to ones commonly found in the relevant literature, since there is no objective ground truth for the task. It was performed on 3 stereoscopic Hollywood movies released in 2011, hereby named "Movie1", "Movie2" and "Movie3". 10 subjects (9 naive and 1 expert) were asked to rate each of the final video skims, in relation to the original movies, in two separate ways: in terms of their *informativeness* and in terms of their *enjoyability*. These two rates per video skim were given on a 0% - 100% scale. All subjects, having recently watched the 3 movies, were independently shown the skims in a consecutive manner and in random order. As in [20], the scale was graded the following way: poor (0% - 40%), fair (40% - 60%), good (60% - 75%), very good (75% - 90%) and excellent (90% - 100%).

In the context of this study, informativeness refers to video content coverage achieved by the produced skim, i.e., to what degree the latter is representative of the original video, retains major plot points and successfully demonstrates major role relationships. Enjoyability refers to the aesthetics of the produced video skim, i.e., to what degree it is composed of semantically complete and coherent scenes, without abrupt and unnatural changes, while simultaneously preserving exciting movie segments and not containing unessential or redundant shots / scenes.

Two main skims were evaluated per movie. One was derived from a summarization pipeline that did not include the MSP step, while the other from one that did apply the proposed shot selection algorithm. The presence of MSP in the second pipeline implies that the corresponding skims are shorter in duration than the ones produced by the first one, allowing us to evaluate the success of the proposed shot pruning scheme. In both cases, p was set to 0.5, precomputed shot cut boundaries were provided and the image modalities used for video frame description were luminance, stereoscopic disparity and color / hue.

TABLE I. A COMPARISON OF THE MEAN INFORMATIVENESS SCORES FOR THE THREE FEATURE FILMS USED IN THE EVALUATION PROCESS.

Method	Movie1	Movie2	Movie3
MSP Pipeline	70%	74%	72%
No-MSP Pipeline	83%	82%	81%

TABLE II. A COMPARISON OF THE MEAN ENJOYABILITY SCORES FOR THE THREE FEATURE FILMS USED IN THE EVALUATION PROCESS.

Method	Movie1	Movie2	Movie3
MSP Pipeline	72%	73%	71%
No-MSP Pipeline	56%	59%	57%

The results of the subjective evaluation are shown in Tables I and II. The MSP Pipeline achieves significantly better enjoyability scores, at the cost of slightly reduced informativeness, which is to be expected since the duration (in total number of frames) of the skims derived through the MSP Pipeline is roughly half that of the corresponding No-MSP Pipeline skims, as it can be seen in Table III. These results suggest that the additional post-processing stage successfully removes redundant movie segments and leads to a skim composed of more complete and coherent scenes, while preserving (at least to a degree) major role relations and plot points.

TABLE III. DURATION (IN FRAMES) OF THE VIDEO SKIMS PER MOVIE. THE DURATION OF THE ENTIRE MOVIE IS ALSO PROVIDED.

Method	Movie1	Movie2	Movie3
MSP Skim	24644	34907	28020
No-MSP Skim	54879	67771	76880
Entire Movie	150358	181763	196224

Regarding the genetic algorithm, the following parameters were used for all movies: the maximum number of generations was set to 200, the population size was set to 200, the crossover rate was set to 0.9, the mutation rate was set to 0.005 and the elitism rate was set to 10%. The number of detected visible speakers for Movie1 was 24, for Movie2 was 13 and for Movie3 was 20, while the corresponding values of C (i.e., the number of movie shots retained after solving the CSSP) are 540 (out of 2161 total detected shots), 511 (out of 2044 shots) and 631 (out of 2524 shots).

The mean required execution time per video frame across all movies, taking into account all pipeline stages, was 857 milliseconds for the No-MSP Pipeline and 1632 milliseconds for the MSP Pipeline. This execution times were measured on a high-end desktop PC, with a Core i7 CPU @ 3.5 GHz and 16 GB RAM. There is an obvious trade-off between summarization quality and execution speed, implying that the proposed method is only suitable for off-line applications.

IV. CONCLUSIONS

We have proposed a multimodal, narrative-preserving shot selection method for movie summarization (Multimodal Shot Pruning, MSP) that does not require extra data modalities beyond the film itself. It tends to only retain key-segments/key-frames belonging to an optimal subset of shots, with subset optimality expressed in terms of discarding shots where only supporting characters are appearing or speaking. The ranking of characters according to the degree they are lead or supporting ones is implicitly and automatically performed by the method, through rudimentary scene modelling. MSP can be integrated into any movie summarization algorithmic pipeline, as a post-processing refinement step. It was evaluated through a standard subjective evaluation process, using three stereoscopic 3D Hollywood movies and a state-of-the-art dynamic movie summarization pipeline, achieving great increases in terms of summary enjoyability at the cost of slight drops in informativeness. The results indicate the importance of intelligent, narrative-conforming shot selection schemes, such as the proposed method.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] Y. Li, S. H. Lee, C. H. Yeh, and C.-C.J. Kuo, "Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 78–89, 2006.

[2] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: state of the art and challenges," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[3] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, 2011.

[4] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proceedings of International Conference on Image Processing (ICIP)*, vol. 1, pp. 866–870, 1998.

[5] S. E. F. De Avilla, A. P. B. Lopes, A. L. Jr. Luz, and A. A. Araujo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

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

[7] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *International Journal on Digital Libraries*, vol. 6, no. 2, pp. 219–232, 2006.

[8] J. Monaco, *How to Read a Film: The Art, Technology, Language, History and Theory of Film and Media*, Oxford University Press, 1982.

[9] J. Sang and C. Xu, "Character-based movie summarization," in *ACM International Conference on Multimedia*, 2010, pp. 855–858.

[10] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," in *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. 2009, SODA '09, pp. 968–977, Society for Industrial and Applied Mathematics.

[11] I. Mademlis, N. Nikolaidis, and I. Pitas, "Stereoscopic video description for key-frame extraction in movie summarization," in *European Signal Processing Conference (EUSIPCO)*. 2015, pp. 819–823, IEEE.

[12] G. N. Stamou, M. Krinidis, N. Nikolaidis, and I. Pitas, "A monocular system for automatic face detection and tracking," *Proceedings of Visual Communications and Image Processing*, vol. 5960, 2006.

[13] O. Zoidi, A. Tefas, N. Nikolaidis, and I. Pitas, "Stereo object tracking with fusion of texture, color and disparity information," *Signal Processing: Image Communication*, vol. 29, no. 5, pp. 573–589, 2014.

[14] G. Orfanidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Facial image clustering in stereoscopic videos using double spectral analysis," *Signal Processing: Image Communication*, vol. 33, pp. 86–105, 2015.

[15] O. Zoidi, A. Tefas, N. Nikolaidis, and I. Pitas, "Person identity label propagation in stereo videos," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1358–1368, 2014.

[16] A. Civril and M. Magdon-Ismail, "Column subset selection via sparse approximation of SVD," *Theoretical Computer Science*, vol. 421, pp. 1 – 14, 2012.

[17] A. K. Farahat, A. Elgohary, A. Ghodsi, and M. S. Kamel, "Greedy column subset selection for large-scale data sets," *Knowledge and Information Systems*, pp. 1–34, 2014.

[18] A. Deshpande and L. Rademacher, "Efficient volume sampling for row/column subset selection," *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 329–338, 2010.

[19] P. Kromer, J. Platos, and V. Snasel, "Genetic algorithm for the column subset selection problem," *Complex, Intelligent and Software Intensive Systems*, pp. 16–22, 2014.

[20] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553 – 1568, 2013.