Location Based Abstraction of User Generated Mobile Videos

Onni Ojutkangas, Johannes Peltola, and Sari Järvinen

VTT Technical Research Centre of Finland
Kaitoväylä 1, 90570 Oulu, Finland
{onni.ojutkangas,johannes.peltola,sari.jarvinen}@vtt.fi

Abstract. Demand for efficient ways to represent vast amount of video data has grown rapidly in recent years. The advances in positioning services have led to new possibilities of combining location information to video content. In this paper we present an automatic video editing system for geotagged mobile videos. In our solution the system creates automatically a video summary from a set of unedited video clips. Geotags are used to group video clips with the same context properties. The groups are used to create a video summary where the videos from the same group are represented as scenes. The novelty in our solution lies in the combining of geotags with low level content analysis tools in video abstraction. Evaluations of the system prove the concept useful as it improves coherence and enjoyability of the automatic video summaries.

Keywords: video summarization, context awareness, video content analysis, temporal segmentation.

1 Introduction

The significance of digital video applications and their market potential has grown remarkably in past years. Drivers for this change can be found from increased processing power of computers and portable devices, larger capacity of storages, faster networks, and digitalization of video. Consumers have moved from using old tape camcorders to digital video recording with devices such as mobile phones and digital cameras. Video recording with mobile phones is fundamentally different than with traditional camcorders. Traditionally videos were recorded at special occasions such as birthdays, festivals and when travelling and the use of the recorder was planned beforehand. Also more effort was given to editing the content. In contrast, as people carry mobile phones with them video recording has become more spontaneous and video is recorded at situations where it previously was not.

The growing amount of personal video content has led to problems with management and consumption of videos. Tools and methods for video editing available to consumers are modeled after those of professional video production, even if consumers usually possess neither time, money, nor expertise that professional production methods require [19]. In order to overcome this obstacle for personal

L. Atzori, J. Delgado, and D. Giusto (Eds.): MOBIMEDIA 2011, LNICST 79, pp. 295–306, 2012.

video management a number of solutions for automatic video editing have been presented. The published work includes systems used for various video domains: sport [3], news [4], documentaries, movies [5], lecture recordings and home videos [1],[2][6],[7]. However, automatic video editing of mobile videos is rarely discussed. In [8] the authors present a video editor for mobile phones based on user studies. They identify the following user goals for video editing in mobile context: select the clips to be edited from the raw source material, combine several separate video clips into one video, cut a clip, enhance the video with text, images, music, and special effects, store the completed video in the device and share the created videos with family or with peer group.

An effective way to make the viewing experience more entertaining is to present the video in as compact form as possible. Hua et al. [2] presents a system, which automatically selects suitable or desirable highlight segments from a set of unedited home videos and aligns them with a given piece of incidental music.

Truong et al. [17] urges more work to be done on maintaining the context and coherence of generated video summaries. Usually shots are joined together simply by their temporal order. However when a video summary is done from a large video collection of material from multiple users from multiple locations, this approach is not feasible.

The targeted domain in our work is personal mobile videos. Techniques proposed for video summarization often use domain specific features, for instance applause or cheering of audience in sport videos. Lienhart [3] uses an empirically motivated approach to cluster time-stamped shots of home video. The approach was encouraged by the fact that the moment of shooting plays a significant role in the home video domain. Lienhart identifies the following unique features for home videos that also apply to personal mobile videos:

- home videos don't have artificial story, plot or structure and
- home video consists of unedited, raw footage.

Our own observations from previous work on mobile videos [18][20][21] revealed that mobile videos include also unique features. Mobile videos are recorded freehanded with small and light devices and this sometimes leads to jerkiness or unstableness in the video. Another observation is that instead of recording several shots average user tends to use panning to cover the surroundings of scenery. Videos recorded with a mobile phone are also often relatively short as the motivation for capturing a mobile video usually is to record a surprising and interesting event or happening for reminiscing or sharing. In addition to this the capturing of mobile videos is not usually planned beforehand unlike use of camcorders. One feature of current video recorders in mobile phones is that they can shoot high quality video in good lighting conditions but they suffer from poor quality in the gloom.

A mobile phone as video recorder provides also means for sensing of context and associating this information to video content. Dey and Abowd [13] [14] regard identity and activity as primary types of context. These primary context types can be used as indices to access secondary context metadata such as a list of friends, user preferences or other people at the same location. This kind of metadata could be used for personalizing the video abstraction.

Geotagging is becoming more popular as integrated GPS modules are found more and more in consumer products such as Nokia N-series multimedia phones, Apple iPhone and Nikon Coolpix P6000 digital camera. Popular multimedia services such as Flickr offer their users various ways to associate location information into their multimedia content. On the other hand map services like Google Maps or Google Earth can be used to visualize the multimedia content using a map interface.

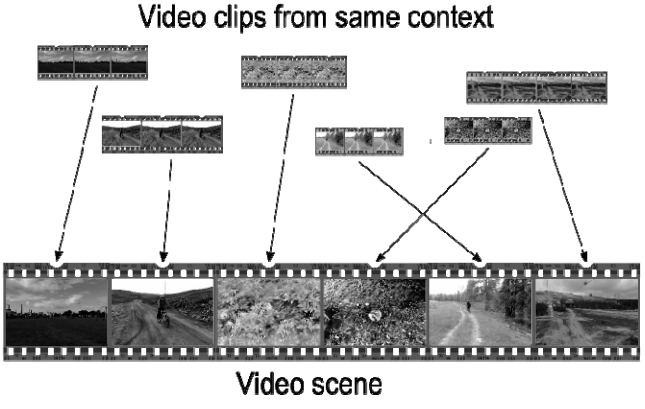

Fig. 1. Concept of a video scene

A novel approach for abstraction of video content is introduced in this work. The goal was to build a system to automatically edit a coherent video summary from a set of unedited user generated mobile video clips created by a group of users. Our aim is to enhance community-based communication by creating an easy-to-use service that allows composing different video summaries based on commonly created video content. Creation location (GPS coordinate) and time information associated with the video clips are used to group shots from same context together to collections of video clips for creation of scenes (Fig. 1). This approach follows the definition [16] of a video scene which depicts action in one location and time. The video clips are segmented and shots to be included in the summary are selected based on results from audio and video analysis taking into consideration parameters such as segmentation threshold, motion intensity weight, motion variance weight and sound weight. A summary of videos is presented as a sequence of scenes. Fading transition effect is used between video scenes to express passage of time and changing of context to the viewer.

The system functionality has been evaluated using a set of video clips acquired from real users of the content creation platform presented in [18]. The video content has been used to generate various video summaries for a small scale user evaluation. The purpose for the evaluation was to get a first impression of the applicability of the chosen approach in enhancing the quality and the amount of information of the video summaries.

The paper is organized as follows; in section 2, we describe the developed video summarization system. In section 3, we present results from an evaluation performed with a small user group. Finally, in sections 4 and 5, we provide concluding remarks and our thoughts of future work.

2 Video Summarization System

Basic steps of video summarization in this system include video segmentation, subshot selection, scene creation and video rendering. Video segmentation is done by analyzing the coherence of video content in frame level. In subshot selection phase a score that reflects the level of interest in a particular subshot is calculated and the most interesting subshots are selected. Video scenes are created by clustering selected subshots by their creation location coordinates and timestamps. Final video summary is rendered by connecting video scenes together and using fading transition effect between them. These steps are described in more detail in the following chapters.

2.1 Feature Extraction

Segmentation of video cannot be done without analyzing the content of video. Video content is analyzed by calculating feature (Table 1) values from audio and video stream. Image brightness, image contrast and similarity features are all calculated from image histograms. Similarity of consecutive frames is measured with match value

$$match(h_1(i), h_2(i)) = \frac{\sum_{j=0}^{L-1} \min\{h_1(j), h_2(j)\}}{\sum_{j=0}^{L-1} h_1(j)}$$

where h_1 and h_2 are compared histograms with L bins. These features are lightweight and give general information of visual content. Camera and object motion are calculated from optical flow of two consecutive frames. Optical flow is calculated with Lucas Kanade method utilising OpenCV library [8]. Root-mean-squared (rms) value of sound signal is calculated over a sequence of n samples, which is set to match the video frame rate:

$$n = \left\lfloor \frac{number\ of\ audio\ samples}{number\ of\ video\ frames} \right\rfloor = \left\lfloor \frac{audio\ sample\ rate}{video\ sample\ rate} \right\rfloor$$

Table 1. Extracted features

Feature	Measure
Image brightness	Mean of image Y channel intensity values
Image contrast	Variance of image Y channel intensity values
Similarity of consecutive frames	Match values of Y, U, and V components
Camera motion	Mean of optical flow vector x and y components
Object motion	Variance of optical flow vector magnitudes
Sound power	Sound RMS power

2.2 Video Segmentation

In video segmentation phase the whole video sequence is segmented to base units, subshots. Low level video parsing can be seen as the analysis of content coherence of a video stream. Subshot boundaries are detected by computing the values of features for each frame of the clip and applying a coherence discontinuity detection mechanism. Coherence of a subshot is analyzed by comparing feature values of two consecutive frames and also comparing average of values before and after the frames. This method is presented in Fig 2.

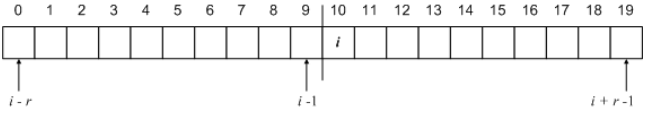

Fig. 2. Segmentation decision is made comparing feature vectors of two consecutive frames, i and i-1, and average vector of r previous frames and r next frames

Similarity of feature vectors is measured with normalized Euclidean distance

$$D(\bar{x},\bar{y}) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}},$$

where x_i and y_i are elements of feature vectors $\bar{x}$ and $\bar{y}$.

Average feature values represented in Table 2 are calculated for each subshot. Personalization of the video summary is accomplished by weight factors a_j given for each feature.

Table 2. Features of a subshot and expected effect of weight factor to the final video summary

Feature of subshot	Expected effect on final summary
Brightness	Giving greater weight to this feature subshots recorded in poor lighting conditions are not likely to be selected to the video summary.
Contrast	Subshots with narrow contrast value are not likely to be included to the video summary if this feature is given weight.
Motion intensity	This feature should be given weight if a video summary with camera movement is desired.
Motion variance	If this feature is given weight scenes with moving objects are expected to be found in the final video summary.
Sound power	By giving weight to this feature subshots including sounds are more likely to be included in the video summary.

2.3 Subshot Selection

In order to create a video summary with the best possible user experience the system tries to preserve the most interesting subshots. The level of interest is modeled with a linear combination

$$\sum_{i=1}^{m} a_i z_i ,$$

where z_i is the subshot feature value and $a_{i.}$ the feature weight.

Selection of the subshots can be seen as a 0-1 knapsack problem: maximizing the total perspective score while not exceeding the upper bound of the summary length. In this work a summarization ratio is known a priori information. Summary length can be easily calculated by multiplying total duration of videos with the summarization ratio. The knapsack problem can be formulated as an optimization problem:

$$\text{maximize} \quad \sum_{j=1}^{n} p_j x_j$$

$$\text{subject to} \quad \sum_{j=1}^{n} w_j x_j \le c, \quad x_j = 0 \text{ or } 1, \quad j = 1, \ldots, n.$$

Here p_j is a profit, w_j is a weight of an item j and c is capacity of the knapsack. In subshot selection phase this problem needs to be solved.

In this work greedy algorithm is used to solve the optimization problem. Algorithm arranges items to descending order according to their p_i/w_i ratio. Then items are added to the sack starting from the first item and stopped when no items fit to the sack anymore. In this case the duration of a subshot is the weight value w_j. The profit value p_j represents the level of interest value of a subshot.

Duration of a subshot cannot be used as subshot feature in this case, because of the greedy algorithm used for solving the subshot selection. This can be simply discovered by assuming duration feature to be z_1, which means that p_i/w_i ratio would be

$$\frac{a_1 z_1 + a_2 z_2 + \cdots + a_m z_m}{z_1} = a_1 + \frac{a_2 z_2 + \cdots + a_m z_m}{z_1},$$

where duration weight value a_1 would have no effect on ordering of subshots.

2.4 Subshot Clustering

After the segment selection the selected segments are organized in two-level hierarchical manner. First the subshots are divided into groups according to their creation location and time information. This approach is following the definition of a scene which depicts action in one location and time.

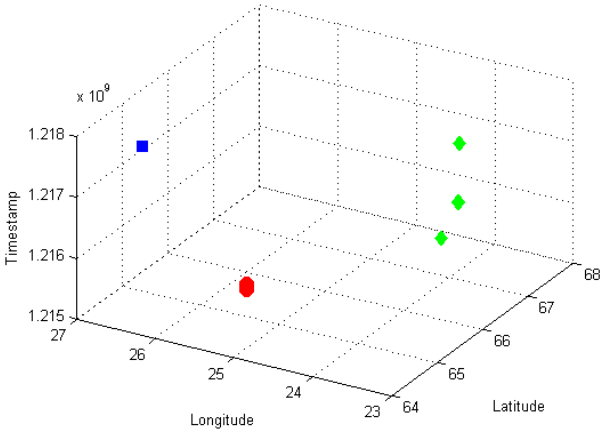

Fig. 3. Result of the clustering algorithm using video creation timestamp and location (latitude and longitude). Subshots from 6 video clips were assigned to three clusters visualized with squares, circles and diamonds.

Subshot grouping is accomplished with an iterative *K*-means clustering algorithm which clusters the subshots in three-dimensional vector space using the creation time and location information (latitude and longitude) of the clip (Fig. 3). The algorithm finds iteratively natural mean points for the clusters and classifies data points to a cluster with nearest mean. Data points are assumed to be defined in a vector space where the chosen metric can be applied. In this work normalized Euclidean distance is used as metric for distance of two vectors.

K-means clustering algorithm assumes the number of clusters K as a priori information. The algorithm tries to minimize a squared error

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (\bar{x}_j - \bar{\mu}_i)^2$$

where $\bar{x}_j$ is data point j and $\bar{\mu}_i$ is mean or center of cluster S_i.

In this work a variation of this algorithm, *K*-medoids, is used. In the algorithm the cluster means are forced to be one of the data points, a medoid.

Performance of the algorithm is heavily influenced by the initialization of cluster means. It is impossible to know beforehand where the clusters are located, after all that is what the algorithm is supposed to solve in the first place. This issue is handled by choosing initial means randomly and running the algorithm several times to find the best clustering.

Another challenge is to choose the right number of clusters K. Approach taken in this work is to iteratively run the algorithm ten times and then increment the value of K. This is repeated until the squared error V goes below a predefined threshold.

2.5 Video Rendering

In video rendering phase the objective is to assemble video scenes to a video sequence and render it to a final video summary. The scenes are ordered chronologically by comparing creation times of the median subshots of clusters. At the second level subshots are ordered inside each cluster by their temporal order. This is particularly important if subshots from the same original video clip are represented in a video scene.

Subshots can be assembled together with various types of transition effects, such as fades, wipes and dissolves. Proper use of shot transition effects can greatly improve the overall appearance of video summary. However some transition effects can have a unwanted effect of drawing attention to the actual effect rather than to the actual transition and therefore the use of effects should be carefully thought over in design of such a system for a specific application. In this work fading transition effect is chosen to be used between video scenes because it generally represents passage of a time or change of context to the user. Simple cuts are used between shots inside a video scene.

Problematic integration of audio modality must be considered when creating video summaries. The problem is that the subshot segmentation is hard to make in a way that would take into account both audio and video modality coherences at the same time. This sometimes leads to awkward cuts in either audio or video stream. Simply combining original audio stream of the subshots to video summary may not be the most appealing solution. In this work a completely unrelated audio stream, for example a music clip, can be added to the summary. This way the video summary should feel more consistent for the viewer.

3 Experiments and Results

In order to have a first impression on the feasibility of the automatic video editing system we organized a small scale user evaluation. In the user evaluations effect of system parameters on experienced *amount of information* and *enjoyability* was measured. *Amount of information* is described as how well the generated summary is considered to represent the whole original video material and *enjoyability* as how pleasing the summary is to watch.

The video collection used in the user evaluations included 48 video clips. Videos were gathered from mobile multimedia content creation platform [18] test conducted in summer 2008. In the tests 7 users got Nokia N95 multimedia phones with context metadata acquiring functionality installed in them and an assignment to shoot and share videos with the content sharing platform. Total video material used in these user evaluations sums up to 11 minutes and 30 seconds.

Five video summaries were created for the user evaluations using the whole video collection as starting point. First summary was generated with random segment selection and assembly. This summary is used as a reference when rating other summaries. Summaries were created with summary ratio of 0.1 so each video is about one minute long summary of total 11 minutes and 30 seconds of video material.

System parameters chosen to be tested were segmentation threshold, motion intensity weight, motion variance weight and sound weight. Parameters used in generation of each video summary are presented in Table 3.

Table 3. Summary parameter values. Parameter t_{segm} is the segmentation threshold, a_{mi} is the motion intensity weight, a_{mv} is the motion variance weight, and a_s is the sound weight.

Summary	$\mathbf{t_{segm}}$	$\mathbf{a_{mi}}$	$\mathbf{a_{mv}}$	$\mathbf{a_s}$
2	2.2	-0.58	-0.11	0.0
3	1.6	-0.58	-0.11	0.0
4	2.2	-0.58	-0.11	30.0
5	2.2	5.8	1.1	0.0

With these four parameters the pace and camera movement of the video summary are changed. Second video summary includes long duration shots with static camera, and fifth summary includes lot of camera movement and shorter shots. Third summary is similar to the second but with shorter shots and thus faster pace. In video summary 4 some different shots were chosen than in the second video summary. Notable difference is that many shots were selected from a football match where sound power was higher than in general shots.

We recruited 8 students of which 7 were men and one woman for these evaluations. Ages of the evaluators were between 20 and 24 years. We asked some details of their video creation and consumption habits in order to have an idea on their expertise in video recording and editing. Six of them said they recorded videos with their mobile phone. Six evaluators also said they recorded videos with digital cameras or digital camcorders. Only three of them were editing their videos afterwards. Software that these evaluators used was Apple iMovie [23], Pinnacle Studio [25], Kino [26] and Windows Movie Maker [27].

The evaluators were first given a short introduction to the subject and the video collection. Then they were asked to view the video summaries from 1 to 5 in consecutive order and rate the experienced level of *amount of information* and *enjoyability* with a scale from 0 to 9 where zero is the lowest level. Values of first video summary were fixed to 4. Results of the evaluations are represented in Table 4.

Table 4. Results from the evaluation of *amount of information / enjoyability*

Summary	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**mean**
1 (ref.)					8/8						4.0/4.0
2				2/1	1/0	0/3	2/2	2/2	1/0		5.3/5.5
3					0/1	2/1	1/1	1/2	4/2	0/1	6.9/6.6
4			2/0	1/2	1/2	1/2	2/1		1/1		4.5/4.8
5			1/0	1/1	0/2	2/3	1/1	1/0	2/1		5.5/5.0

All created summaries were generally considered to be more informative and enjoyable than the summary with random segment selection and assembly. Results show that video summary three was experienced to be the most informative and enjoyable. In summary 4 the *amount of information* and *enjoyability* probably suffered from the multitude of football scenery. Summaries with more camera movement were generally considered less enjoyable than summaries that included shots with static camera. However the experienced *amount of information* did not seem to be lower in summary with more camera movement.

4 Conclusion

Video abstraction tools can be effectively used to visualize large amount of video content, such as user generated mobile videos. The advances in positioning services have led to new possibilities of combining location information to video content and use it for content management purposes.

In this paper we introduced a novel approach to utilize location and time metadata in video abstraction together with more traditional content analysis tools. Video clips are segmented to subshots by analyzing video content in frame level. A level of interest score is calculated for each subshot and the most interesting subshots are selected for the summary. The location and time related metadata is used to cluster subshots from the same context together to form video scenes. These scenes are rendered to a video summary by using fading transition effect between scenes and hard cuts between shots inside a same scene.

This approach was found to be very useful as the coherence and *enjoyability* of generated video summary was greatly enhanced based on our small scale user evaluation. In these evaluations the generated video summaries were considered to be more enjoyable and informative than summaries generated with random subshot selection and assembly. Video summaries that included static camera were found to be the most enjoyable.

5 Future Work

The performed user evaluations showed that the idea of automatic video editing system based on geotagging and timestamps of the original video clips is a feasible idea. From the user perspective it is essential to define the correct ratio of automation and manual work in creation of the video summaries. We need to study further the effect of system parameters in the final summary taking into account the characteristics of mobile video clips in order to be able to provide best possible service and user interface for the end user to create his or her personal and community video summaries. The result of this work will need to be examined with a larger scale user test to validate the functionality and usability of overall automatic video editing system and user interface.

There is also a lot to do on algorithm development to ensure the scalability of our system. Scalability requirements can be met by optimizing content analysis and

encoding algorithms as well as distributing processing e.g. in a cloud environment. The results from various audiovisual content analysis methods can be integrated to the automatic video editing system in order to add more features to the content clustering process. In order to provide the best possible summaries for specific use cases the logic for summary generation should be designed separately for each application domain. Adding an audio track to the background will surely increase the *enjoyability* of the video summary.

Acknowledgements. This work was part of ITEA2 project ExpeShare – Experience Sharing in Mobile Peer Communities (2007-2009) partially funded by TEKES the National Technology Agency of Finland.

References

1. Hua, X.-S., Lu, L., Zhang, H.-J.: Optimization-Based Automated Home Video Editing System. IEEE Transactions on Circuits and Systems for Video Technology 14(5) (2004)
2. Mei, T., Hua, X.-S., Zhu, C.-Z., Zhou, H.-Q., Li, S.: Home Video Visual Quality Assesment With Spatiotemporal Factors. IEEE Transactions on Circuits and Systems for Video Technology 17(6) (2007)
3. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Transactions on Image Processing 12(7) (2003)
4. Huang, C.-L., Hsieh, C.-H., Wu, C.-H.: Audio-video summarization of TV news using speech recognition and shot change detection. In: 9th European Conference on Speech Communication and Technology, International Speech and Communication Association (2005)
5. Evangelopoulos, G., Rapantzikos, K., Potamianos, A., Maragos, P., Zlatintsi, A., Avrithis, Y.: Movie summarization based on audiovisual saliency detection. In: 15th IEEE International Conference on Image Processing (2008)
6. Lienhart, R.: Abstracting home video automatically. In: Proceedings of the 7th ACM International Conference on Multimedia, Part 2 (1999)
7. Zhao, M., Bu, J., Chen, C.: Audio and video combined for home video abstraction. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003 (2003)
8. Jokela, T., Karukka, M., Mäkelä, K.: Mobile Video Editor: Design and Evaluation. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4551, pp. 344–353. Springer, Heidelberg (2007)
9. OpenCV documentation of optical flow algorithms OpenCV Wiki, `http://opencv.willowgarage.com/wiki/CvReference#OpticalFlow` (accessed April 9, 2009)
10. Yu, B., Ma, W.-Y., Nahrstedt, K., Zhang, H.-J.: Video summarization based on user log enhanced link analysis. In: Proceedings of the Eleventh ACM International Conference on Multimedia (2003)
11. Adami, N., Benini, S., Leonardi, R.: An overview of video shot clustering and summarization techniques for mobile applications. In: Proceedings of the 2nd International Conference on Mobile Multimedia Communications (2006)
12. Huet, B., Merialdo, B.: Automatic Video Summarization. In: Interactive Video, Signals and Communication Technology. Springer, Berlin (2006)

13. Dey, A.K.: Understanding and Using Context. Personal and Ubiquitous Computing 5(1) (2001)
14. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a Better Understanding of Context and Context-Awareness. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 304–307. Springer, Heidelberg (1999)
15. Davis, M., King, S., Good, N., Sarvas, S.: From context to content: leveraging context to infer media metadata. In: Proceedings of the 12th Annual ACM International Conference on Multimedia (2004)
16. Goodman, R.M., McGrath, P.: Editing Digital Video: The Complete Creative and Technical Guide, 1st edn. McGraw-Hill, Inc. (2002)
17. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM Transaction on Multimedia Computing, Communications, and Applications 3(1) (2007)
18. Järvinen, S., Peltola, J., Plomp, J., Ojutkangas, O., Heino, I., Lahti, J., Heinilä, J.: Deploying mobile multimedia services for everyday experience sharing. In: Proceedings of IEEE International Conference on Multimedia and Expo., ICME 2009 (2009)
19. Davis, M.: Editing out video editing. IEEE Multimedia 10(2) (2003)
20. Pietarila, P., Westermann, U., Jarvinen, S., Korva, J., Lahti, J., Lothman, H.: CANDELA – storage, analysis, and retrieval of video content in distributed systems. In: Proceedings of the IEEE International Conference on Multimedia and Expo. (ICME 2005), pp. 1557–1560 (2005)
21. Järvinen, S., Peltola, J., Lahti, J., Sachinopoulou, A.: Multimedia service creation platform for mobile experience sharing. In: Proceedings of the 8th International Conference on Mobile and Ubiquitous Multimedia, MUM (2009)