A Survey on Multimodal Video Representation for Semantic Retrieval

Janko Ćalić *, Neill Campbell *, Stamatia Dasiopoulou † and Yiannis Kompatsiaris †

Abstract - **This paper surveys the approaches to video representation, focusing on semantic analysis for content-based indexing and retrieval. A problem of adaptive representation of digital multimedia is critically assessed and some novel ideas are presented. Furthermore, the concept of video multimodality is reevaluated and redefined in order to introduce modalities such as editing technique or affect to the audience.**

Keywords - **video representation, multimodality, content-based indexing and retrieval, semantic gap**

I. Introduction

In the field of multimedia content-based retrieval there has been a plethora of interesting research work presented recently that focuses on problem of the *semantic gap* between low-level information extracted from the media and the user's need to meaningfully interact with it on a higher level. However, the majority of ideas follow a paradigm of finding a direct mapping from low-level features to high-level semantic concepts. Not only does this approach requires extremely complex and unstable computation and processing [3], but it appears to be unfeasible unless it targets a specific and contextually narrow domain [25, 4, 15]. Little has been done to design a system capable of creating appropriate representations of video media on various levels of complexity and thus improve adaptability and reliability of multimedia retrieval. As given in [2, 27], the multimedia database stands as a central point of the modern creativity and thus the challenge to effortlessly interact with the large digital media collections is our prime goal. In addition, much of the recent research attempts to utilise the *multimodal* character of the video media [41, 26, 35] but fails to fully exploit the underpinning information from these closely intertwined modalities.

This paper is an attempt to shed new light on multimodal video representation and to give a critical survey of publications in this field. The everlasting question of the optimal representation of digital video media, targeting not only content-based retrieval but content analysis in general, is reassessed in order to identify the reasons behind such a persistent problem as the sematic gap. There have been numerous literature surveys on the various aspects to content-based video indexing and retrieval [37, 38, 41]. However, the problem of appropriate multimodal representation has been practically ignored, since the choice of low-level features in current retrieval systems tends to be independent of the content and its semantics.

Following a similar approach to the problem of video representation, the concept of video media multimodality is critically rethought outlining the outstanding approaches that challenge this problem. To support these claims, a survey of two common approaches to multimodal video representation, opposite in their character, is given i.e. data driven and concept driven generation of representation models.

The structure of the paper is as follows. The following section outlines the problem of video representation. The section III brings a novel perspective to the concept of multimodality in digital video media. Data driven approaches to representation is given in section IV and top-down algorithms are presented in section V. The paper finalises with major conclusions and the list of the referred publications.

II. Video representation

In the context of semantic retrieval of video media, we address the problem of *computational* video representation, i.e. how to abstract the audio-visual experience of the user by means of computational models. This is clearly a difficult task that has to involve both appropriate computation and processing as well as the way in which a user experiences targeted media. However, this is not a common approach to video retrieval, where the focus is on the way information is extracted from the digital media, whether it makes sense to the user or not.

The foundational work that has formulated the problem of computational video representation was presented by Davis [13, 12] and Davenport et al. [11]. In [12] multi-layered, iconic annotations of video content called MediaStreams is developed as a visual language and a stream-based representation of video data, with special attention to the issue of creating a global, reusable video archive. Being radically oriented towards a *cinematic* perspective of video representation, the work presented in [11] sets the scene to a novel approach to the content-based video analysis based upon a shot, an irreducible constituent of video sequences. But this was where research community stopped following this paradigm and got attracted to extraction and analysis of low-level features, ignoring the fact that these features would make little or no sense to the end user.

And it wasn't until the definition of Computational Media Aesthetics (CMA) in a number of publications by Dorai and Venkatesh [17, 16] that the user centered representation reemerged within the video retrieval community. The

[*] Janko Ćalić and Neill Campbell are with the Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK, e-mail: {janko, campbell}@cs.bris.ac.uk

[†] Stamatia Dasiopoulou and Yiannis Kompatsiaris are with the Informatics and Telematics Institute, 1st Km Thermi-Panorama Road, Thessaloniki, GR-57001 Greece, e-mail: {dasiop, ikom}@iti.gr

main idea behind CMA is to have a focus on domain distinctives, the elements of a given domain that shape its borders and define its essence (in film, for example, shot, scene, setting, composition, or protagonist), particularly the expressive techniques used by a domains content creators [1]. This is clearly a diametrically opposite point of view to the common perception that the video should be indexed by the terms for which automatic detectors can be realized [38]. Nevertheless, these two different approaches are bound to merge in order to achieve the goal of semantic retrieval of video media. Sections IV and V outline work following these approaches, but let us first address the concept of multimodality in video media.

III. Multimodality

Seen from the generic system-centered perspective, *multimodality* is the capacity of the system to communicate with a user along different types of communication channels and to extract and convey meaning automatically [31]. However, the prevailing opinion is that the multimodality of video media is the capacity of an author to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels, where the channels can be either *visual*, *auditory* or *textual* [38]. Is it really true that these are the only communication channels by which the meaning of video is conveyed?

In recent publications challenging fundamental issues in content-based multimedia retrieval [34, 39, 29] there has been an evident turn towards semiotic approach to the problem of the *semantic gap*. Semiotics is the study of signs and of the way meaning is transmitted and understood [36]. It has been applied widely in the analysis of film and video media, underlining the importance of communication through modes such as editing, narrative structure, visual composition (mise en scène), etc. The groundwork for the semiotic analysis in film theory was set by Eisenstein [18], Kuleshov [23] and later by Metz [28]. The *Kuleshov experiment* [19] indicated the importance and effectiveness of film editing by showing that juxtaposing two unrelated images could convey a separate meaning. Given the definition of multimodality [31] stated above, this implies that *editing* is a valid modality of video media. In this manner one could come up with more modalities that could be assigned to video media and ergo the widespread approach to multimodality as merging only visual, auditory and textual information seems to be hindering the development of the semantic analysis in content-based video retrieval.

However, a number of publications show a tendency towards a more sophisticated take on multimodality. The intensity and type of feeling or emotion, both referred to as *affect*, that are expected to arise in the user while watching a video clip have been computationally represented and modelled by Hanjalic and Xu [20]. In the recent publication by Dimitrova [14] an interesting proposition to exploiting concepts of short- and long-term memory in content analysis is presented. In addition, the influence between the multimedia modalities can be modelled through cross-modal association, as presented in the work by Li et al. [26].

Nevertheless, a clear definition of an infinite set of modalities is essential in order to make automatic classification feasible. On the other hand, the number of low-level features extracted from various modalities is limited. Therefore, a balanced interaction between feature driven bottom-up approaches and top-down algorithms could be a solution to a generic model for video representation.

IV. Data driven representations

This approach is the standard way of extracting low-level features and deriving the corresponding representations without any prior knowledge of the related domain. Therefore, this is the hard-encoded way of representation driven by expert knowledge. A rough categorization of data-driven approaches in the literature yields two main classes [42]. The first class focuses mainly on signal-domain features, such as color histograms, shapes, textures, which characterize the low-level audiovisual content. The second class concerns annotation-based approaches which use free-text, attribute or keyword annotations to represent the content.

Many content-based indexing and retrieval systems have been proposed focusing mainly on the definition of suitable descriptors, the generation of appropriate metrics in the descriptors space and efficient addressing of the large workload and high complexity of the underlying image-processing algorithms. The existing systems fall broadly under four categories depending on the chosen content and indexing structures used: query by content, iconic query, SQL query and mixed queries. Query by content is based on images, tabular form, similarity retrieval (rough sketches) or by component features (shape, color, texture). The iconic query represents data with icons as its visual abstraction and specifies a query by the selection of appropriate icons. SQL queries are based on keywords, with the keywords being conjoined with the relationship (AND, OR) between them, thus forming compound strings. The mixed queries can be specified by text as well as icons. For a larger overview of relevant approaches a number of extensive surveys are available [37, 38, 41].

Although useful for representation of video within a limited domain, such approaches lack the capability to adapt to different domains. Indeed, the richness of audiovisual content is difficult to describe with a few keywords and data-driven representations, while content perception itself is a subjective and task-dependent process. The problem is exacerbated by motion and other temporal features. Trying to foresee which elements will be the most useful for subsequent retrieval is very difficult.

V. Top-down approaches

Top-down retrieval systems utilise *high-level* knowledge of the particular domain to generate appropriate representations. This knowledge can be predefined by expert users, semi-automatically learned or acquired in a completely automatic manner. These categories can differ in the way by which high-level information influences extraction and processing of low-level features. In other words, the system can eliminate unimportant features from the initial feature set, re-processes the features in order to generate new representations or influence the feature generation and analysis at the very stage of extraction and processing.

The majority of existing retrieval systems utilise expert knowledge in a hard-coded manner. The representations are generated in a predetermined way and they are not influenced by high-level information of the analysed media [37].

Techniques such as relevance feedback [33] [40] [43] and incremental machine learning [10] enable intervention of the user in the process of knowledge acquisition. There are many promising examples of semi-automated or semi-supervised video retrieval systems that exploit this idea. An approach by Dorado et al. [15] generates the concept lexicon that may consist of words, icons, or any set of symbols that convey the meaning to the user by utilising fuzzy logic and rule mining techniques to approximate human-like reasoning. A nice example of a domain driven semi-automated algorithm for semantic annotation is given by Burghardt [6] where a specific animal face tracker is formed from user labelled examples utilising Ada-boost classifier and Kanade-Lucas-Tomasi tracker.

In the work by Bloehdorn et al.[5] a M-OntoMat-Annotizer is designed in order to construct ontologies that include prototypical instances of high-level domain concepts together with a formal specification of corresponding visual descriptors. Thus, it formalizes the interrelationship of high- and low-level multimedia concept descriptions allowing for new kinds of multimedia content analysis and reasoning.

In a recent overview on supervision and statistical learning for semantic multimedia analysis, Naphade [30] outlines that the problem of small sample statistics limits using traditional learning techniques. However, innovations such as labeled and unlabeled learning, active learning and discriminant techniques have made it more feasible to use statistical models for more general video indexing problems. In their earlier work, Naphade et al.[32] defined a factor graph network of probabilistic multimedia objects, *multijects*, in a probabilistic pattern recognition fashion using hidden Markov and Gaussian mixture models. Another approach that attempts to link a subset of low-level features and words was taken by Barnard et al. in [3] where the joint distribution of image regions and words was learned utilising multi-modal and correspondence extensions to Hofmanns hierarchical clustering/aspect model, a translation model adapted from statistical machine translation and a multi-modal extension to latent dirichlet allocation.

An interesting approach of *emergent semantics* brings a novel way to create meaning within an analysed collection. Emergent computation, as presented by Staab [39], is based on the idea that appropriate semantic structures might arise purely from the physics of the task environment, rather than from an experts elaborate considerations. Specifically focusing on the image databases, Santini [34] claims that images don't have an intrinsic meaning, but that they are endowed with a meaning by placing them in the context of other images and by the user interaction. Since semantics do shape the representation model, a straightforward implication of this idea is that as well as semantics, video representation depends upon the dynamic of the database and interaction with the user, and cannot be predetermined. The idea of emergent semantics could have far-reaching repercussions in the way that content-based retrieval develops.

There are number of domain specific systems that exploit a unique set of features to form a reliable representation. Such examples include semantic analysis of sports, documentaries, newscasts or soap operas. An algorithm presented by Leonardi et al. in [24, 25] exploits hidden markov models on multimodal data to achieve structural and semantic classification of football videos. On the other hand Bertini and Del Bimbo [4] designed a solution for highlights detection in sports videos using finite state machines that encode the temporal evolution of the analysed highlights.

Bearing in mind that the development and evaluation of such a complex task as multimodal video representation requires a large scale content-management framework, there are number of research projects and initiatives that are addressing this problem.

The target of the aceMedia project [22] is the integration of knowledge and multimedia content technologies, focusing on the benefits of the end user, in the context of a user-centered scenario. In order to simplify the user experience, the aceMedia project focuses its efforts on knowledge discovery and self-adaptability embedded into media content, which will allow it to be self organizing, self annotating, and more readily searched and communicated, by providing tools to automatically analyze content, generate metadata and annotation, and support intelligent content search and retrieval services.

The ICBR (Intelligent Content-based Retrieval) system [8] exploits a unique opportunity to deal with semantic gap issues by integrating a large video database with its semantic description organised in a structured taxonomy. This framework proved to be a unprecedented environment for development of novel representation for semantic video analysis of wildlife documentaries [21, 9, 7].

VI. Conclusion

As elaborated above, there is a need for more focus on novelties in video representation in order to tackle the problem of semantic gap. In the situation where the type of the index describing a unit of media is defined by descriptors proposed in the MPEG-7 standard and is limited by the set of index terms for which automatic detectors can be realized [38], the prospect of solving the problem of the sematic gap seems rather remote. Introducing a more sophisticated way of representing information embedded in the video media by interacting with the user and analysing more modalities could bring that essential advance to content-based video indexing and retrieval.

Future work will be focused on defining a generic framework for indexing and retrieval of video in order to evaluate different algorithms for video representation and assessing this novel area in a more objective way.

Acknowledgements

The work reported in this paper has formed part of the activity of the WG3 within IST COST292 action in semantic multimodal analysis of digital media whose funding and support is gratefully acknowledged. For more information please visit www.cost292.org.

REFERENCES

[1] B.D. Adams. Where does computational media aesthetics fit? *IEEE Multimedia Magazine, spec. ed. Computational Media Aesthetics*, April-June 2003.

[2] Steve Anderson. Select and combine: The rise of database narratives. *Res Magazine*, 7(1):52–53, Jan/Feb 2004.

[3] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[4] M. Bertini, A. Del Bimbo, and W. Nunziati. Highlights modeling and detection in sports videos. *Pattern Analysis and Applications*, 2005.

[5] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, I. Kompatsiaris, S. Staab, and M. G. Strintzis. Semantic annotation of images and videos for multimedia analysis. In *Proceedings of the 2nd European Semantic Web Conference, ESWC 2005, Heraklion, Greece*, May 2005.

[6] Tilo Burghardt, Janko Calic, and Barry Thomas. Tracking animals in wildlife videos using face detection. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, October 2004.

[7] Tilo Burghardt, Barry Thomas, Peter J Barham, and Janko Calic. Automated visual recognition of individual african penguins. In *Fifth International Penguin Conference, Ushuaia, Tierra del Fuego, Argentina*, September 2004.

[8] Janko Calic, Neill Campbell, Majid Mirmehdi, Barry Thomas, Ron Laborde, Sarah Porter, and Nishan Canagarajah. ICBR - multimedia management system for intelligent content based retrieval. In *International Conference on Image and Video Retrieval CIVR 2004*, pages 601–609. Springer LNCS 3115, July 2004.

[9] Janko Calic and Barry Thomas. Spatial analysis in key-frame extraction using video segmentation. In *Workshop on Image Analysis for Multimedia Interactive Services*, April 2004.

[10] Gert Cauwenberghs and Tomaso Poggio. Incremental and decremental support vector machine learning. In *Proc. of Neural Information Processing Systems (NIPS) 2000, Denver, USA*.

[11] Glorianna Davenport, Thomas Aguirre Smith, and Natalio Pincever. Cinematic primitives for multimedia. *IEEE Comput. Graph. Appl.*, 11(4):67–74, 1991.

[12] Marc Davis. Media streams: representing video for retrieval and repurposing. In *MULTIMEDIA '94: Proceedings of the second ACM international conference on Multimedia*, pages 478–479, New York, NY, USA, 1994. ACM Press.

[13] Marc Davis. *Media streams: representing video for retrieval and repurposing*. PhD thesis, Cambridge, MA, USA, 1995.

[14] Nevenka Dimitrova. Context and memory in multimedia content analysis. *IEEE MultiMedia*, 11(3):7–11, 2004.

[15] A. Dorado, J. Calic, and E. Izquierdo. A rule-based video annotation system. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):622–633, May 2004.

[16] Chitra Dorai and Svetha Venkatesh. *Media computing: computational media aesthetics*. The Kluwer international series in video computing. Kluwer Academic Publishers, Boston; London, 2002.

[17] Venkatesh S. Dorai, C. Bridging the semantic gap with computational media aesthetics. *Multimedia, IEEE*, 10:15–17, 2003.

[18] Sergei Eisenstein and Jay Leyda. *Film Form Essays in film theory*. Dennis Dobson Ltd, [S.l.], 1949. edited and translated by Jay Leyda.

[19] Louis D. Giannetti. *Understanding movies*. Prentice Hall; London: Prentice-Hall International (UK), Upper Saddle River, N.J., 9th ed. edition, 2002.

[20] A. Hanjalic and L. Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[21] Sion Hannuna, Neill Campbell, and David Gibson. Segmenting quadruped gait patterns from wildlife video. In *The IEE International Conference on Visual Information Engineering: Convergence in Graphics and Vision.*, pages 235–243. Institution of Electrical Engineers, April 2005.

[22] I. Kompatsiaris, Y. Avrithis, P. Hobson, T. May, and J. Tromp. Achieving Integration of Knowledge and Content Technologies: The AceMedia Project. In *Proc. European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, Royal Statistical Society*, London, UK, Nov. 2004.

[23] Lev Kuleshov and Ronald Levaco. *Kuleshov on film: writings by Lev Kuleshov*. University of California Press, Berkeley; London, 1974.

[24] R. Leonardi and P. Migliorati. Semantic indexing of multimedia documents. *Multimedia, IEEE*, 9(2):44, 2002.

[25] R. Leonardi, P. Migliorati, and M. Prandini. Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):634, 2004.

[26] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. Multimedia content processing through cross-modal association. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 604–611, New York, NY, USA, 2003. ACM Press.

[27] Lev Manovich. *The language of new media*. Leonardo. MIT Press, Cambridge, Mass.; London, 2001.

[28] Christian Metz. *[Essais sur la signification au cinéma.] Film language. A semiotics of the cinema. Translated by Michael Taylor*. New York: Oxford University Press, 1974.

[29] F. Nack and A. Parkes. Toward the automated editing of theme-oriented video sequences. *Applied Artificial Intelligence*, 11(4):331–366, 1997.

[30] Milind R. Naphade. On supervision and statistical learning for semantic multimedia analysis. *Journal of Visual Communication and Image Representation*, 15(3):348–369, 2004.

[31] Laurence Nigay and Joëlle Coutaz. A design space for multimodal systems: concurrent processing and data fusion. In *CHI '93: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 172–178, New York, NY, USA, 1993. ACM Press.

[32] M. Ramesh Naphade, I. V. Kozintsev, and T. S. Huang. Factor graph framework for semantic video indexing. *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(1):40, 2002.

[33] J.J. Rocchio, Jr. *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Relevance Feedback in Information Retrieval, pages 313–323. Prentice-Hall, 1971.

[34] S. Santini, A. Gupta, and R. Jain. Emergent semantics through interaction in image databases. *Knowledge and Data Engineering, IEEE Transactions on*, 13(3):337, 2001.

[35] C. Saraceno and R. Leonardi. Indexing audiovisual databases through joint audio and video processing. *IJIST*, 9(5):320–331, 1999.

[36] Ferdinand de Saussure. *Course in general linguistics*. Duckworth, London, 1983.

[37] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[38] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[39] S. Staab. Emergent semantics. *IEEE Intelligent Systems*, 17(1):78–81, 2002.

[40] M. E. J. Wood, N. W. Campbell, and B. T. Thomas. Iterative refinement by relevance feedback in content-based digi tal image retrieval. In *ACM Multimedia 98*, pages 13–20. ACM, September 1998.

[41] Wang Yao, Liu Zhu, and Huang Jin-Cheng. Multimedia content analysis-using both audio and visual clues. *Signal Processing Magazine, IEEE*, 17(6):12, 2000.

[42] A. Yoshitaka and T. Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81–93, Jan/Feb 1999.

[43] Xiang Sean Zhou and Thomas S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst.*, 8(6):536–544, 2003.