A Multimedia Content Semantics Extraction Framework for Enhanced Social Interaction

Giuseppe Passino, Tomas Piatrik, Ioannis Patras and Ebroul Izquierdo
Queen Mary, University of London
Mile End Road, London, E1 4NS, UK
<name>.<surname>@elec.qmul.ac.uk

ABSTRACT

In this paper, a system for improved social interaction via the Internet or interactive TV is proposed. The aim is to provide a small group of closely connected users with a rich social experience, sharing intimate moments of life and emotions, taking full advantage of the existent Internet technology and broadcasting practices. Starting from a single use-case, a feasibility study for a social interaction is illustrated. The proposed architecture for social interaction is based on techniques for automated extraction of semantics from streamed content. In particular, technical feasibility and real-time processing issues are considered. Semantic information is used in the multimedia editing and composition phase, enabling the system to offer an experience that goes beyond the classical face-to-face video-conference. The efficient and rich presentation of the content is driven by technology for semantic segmentation, object detection and automated extraction of interesting regions in the scene. Furthermore, a face detection module is used to guarantee a constant visual presence of the parties. Finally, a summary of the session is automatically generated for future uses or on-line-browsing during the conversation.

1. INTRODUCTION

The pervasiveness of Internet connectivity is generating a social revolution. Users are experiencing deeper interaction independently from remoteness and geographical locations, and it is expected that the use of such remote interactions will continue to grow in the years to come. In this paper we argue that it is possible to integrate classical content-delivery technologies as broadcasting, optimised over many decades, with innovative computer vision techniques to produce a socially rich communication environment through semantics-aware systems, *i.e.*, systems able to react according to the semantics of the streamed content. We present a proposal for an integrated system to enable a rich user interaction experience based on existent broadcasting technology and advanced semantic multimedia analysis.

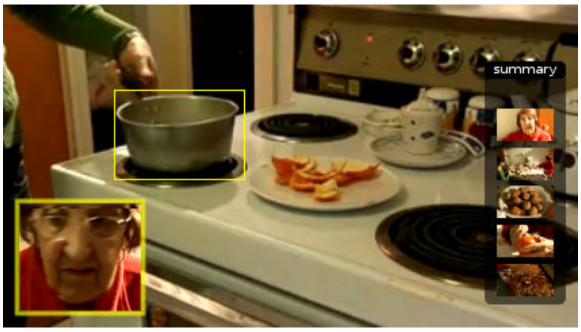

Figure 1: Output video example for the considered use-case.

The scenario considered in this work entails rich social experience between few closely related people, in different locations. The communication is carried out through a multimedia channel, exchanging content acquired through multiple video inputs, audio and complementary user sources. The goal is an immersive experience that goes beyond the classical face-to-face videoconferencing. To illustrate the underlying idea in an effective way the following scenario is used: a young student living abroad is interacting with his mother at home regarding a cake recipe. Using one or more conventional video cameras, the mother can share the preparation of the cake from her own kitchen while communicating in a friendly and close atmosphere, and the content is offered to the son using advanced automated video editing and presentation techniques typical of television, such as zoom and framing effects. The face of the participant may be always visible in a corner of the display, being the face automatically detected and framed by the camera. Kitchen tools, and other interest objects in the scene will be isolated and focused via real-time object detection algorithms. Finally, an automatic summary of the entire video is produced for efficient and easy on-line fruition and later viewing. In Fig. 1 a possible output example is shown.

The remainder of the paper is structured as follows. In Section 2 we present an overview of the proposed system; in Section 3 the face detection problem is presented, while the general and diverse problem of interest object detection is discussed in Section 4. Scene detection is treated in Section 5, and automatic summarisation in Section 6. Expected performance of the components and conclusions are presented in Section 7.

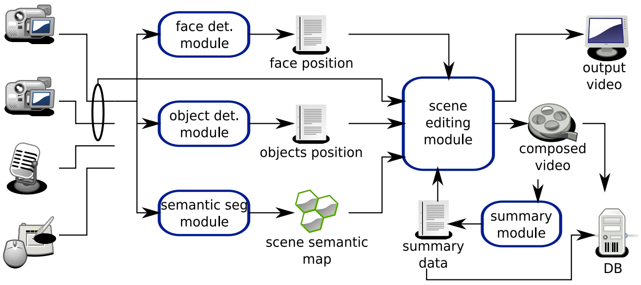

Figure 2: Block diagram of the framework as detailed in Section 2.

2. SYSTEM OVERVIEW

In this work we explicitly address the multimedia processing activities related to semantics extraction, while other aspects of the system are omitted for the sake of conciseness. We do not elaborate on the module for automated content editing and presentation, that should implement techniques borrowed from the TV production world. The basic system modules related to semantics extraction are a face detector, a generic object-of-interest detector, a scene semantic segmentation module, and an automatic summarisation module, as depicted in Fig. 2.

Eye-contact plays a crucial role in communication, for effectively conveying messages and sharing feelings: the face detector component is aimed at spotting the position of the face of the parties, enabling its zooming, framing and constant presentation in a complex final layout. On the other hand, in order for the editing system to react to the current activity, awareness of the objects present in the scene is needed. This requirement is satisfied with a generic object detection algorithm, independent on the particular nature of the object of interest. Finally, a more semantically aware component aimed at the understanding of the entire scene is used to link concepts related to the discussion and the content of the scene (for example, in a kitchen environment, the location of people, cupboards, oven and other appliances can be relevant to the discussion). The content is composed by an automatic content editing component which receives as an input the audio-video streams and the associated semantic information extracted by the detectors. Finally, a summarisation module, applied to the output of the editing block, will automatically produce a set of highlights to easily browse the content, either on-line during the conversation, or for a later usage (for example, to revise the steps of the cake preparation).

3. FACE DETECTION

In the analysed scenario the face detection is not particularly challenging, involving a restricted number of people in an environment with low dynamics. Different optimised algorithms have been used in end-market products, even for portable devices as digital cameras. An approach based on boosting [9] allows for real-time detection. To improve location performances a facial model [1] can be used. A probabilistic model based on a fixed number of stable points is created. The model is then considered as a flexible structure that is efficiently positioned in the best location in the vicinity of the location for the detected face obtained through the boosting-based detector. The detected faces are then tracked with a standard method [6].

4. OBJECT DETECTION

The object detection module is aimed at the identification of the object of interest, in order to focus on it during the video editing stage. This is the most challenging aspect of the framework, since the nature of the objects is unknown by definition, and therefore the object detection module has to be generic: an explicit search for a large set of object categories is not viable in real-time, and its training demanding. For the previously stated reasons, we propose the use of a semantic latent method, in which the categories are not explicitly specified but are instead hidden. We base our approach on the success of the bag-of-words model to detect and localise objects in images [7]. This method is based on probabilistic Latent Semantic Analysis (pLSA), where (visual) words are associated to latent object categories. The categories are therefore described in terms of word histograms. The visual words are salient points in the image encoded in a dictionary of features. The “quality” of the words, as the robustness to image geometric transformation and descriptivity, is affected by the real-time constraint, but the extraction step is generally feasible within these limitations. The adaptive (unsupervised) training step can be done in background and updated at regular intervals.

The output of the pLSA is a set of probabilities of words given the latent categories. From this information we can estimate the category posteriors given the visual words. This is a weak information on the presence and location of an object, as local concentration of words related to a single topic indicate the presence of an object of interest. Since salient points are robust against geometrical transformations, a tracking strategy [6] can be applied to estimate the degree of movement of different objects in the scene.

5. SEMANTIC SCENE SEGMENTATION

The semantic segmentation module is demanded to give an interpretation of the layout of the scene, associating a semantic category label to each image pixel. We propose to use a method developed in our research group, in which the image is analysed by extracting parts and considering their aspect and context [4], with some modifications to respect the real-time constraint required by the application.

The basic module activity can be summarised as follows: image parts are extracted, and descriptors are evaluated for each part; an optimised structure is then computed, to connect the parts; finally, based on the calculated structure and the part features, the parts are labelled through the application of a Conditional Random Field (CRF), a probabilistic graphical model.

This this scenario, since accurate border detection for object is not necessary, a fast part extraction strategy based on a grid of rectangular patches is proposed. The descriptors, based on textons [2] and hue histograms [8], rely on visual dictionaries, that are expensive to compute but need to be compiled offline only once. The optimal structure evaluation is based on the Aspect-Coherent Minimum Spanning Tree (AC-MST), whose associated complexity is quasi-linear, being a MST algorithm. The discriminative nature of the CRF

Table 1: Illustrative pixel-level labelling performance for the semantic segmentation module.

Building	Grass	Tree	Cow	Sky	Airplane	Face	Car	Bicycle	Average
63.0%	94.2%	68.9%	84.4%	93.7%	75.8%	92.9%	76.4%	86.5%	82.9%

makes the inference straightforward in terms of time performance, but the model needs to be trained with pixel-level ground truth, in a time-consuming process. This can be done on-line when installing the cameras, with the help of the user to provide partial image annotation, which is supported by the method.

6. AUTOMATIC SUMMARISATION

The process of creating video summary includes three important steps, namely shot boundary detection, key-frame extraction and shot relevance/redundancy estimation. Many efficient techniques have been proposed to deal with the aforementioned tasks, however most of them are not dedicated for real-time applications because of their high complexity [10, 3]. To reduce the computational complexity as required in this application scenario we propose an approach which combines both scene detection and key-frame extraction to create the summary excluding redundant segments.

A fundamental step in our approach is to create the similarity matrix and organise video frames into a tree structure using Ant-Tree Strategy (ATS) [5]. ATS is inspired by self-assembling behaviour of ants and their ability to build mechanical structures. On the basis of a root on which the tree is built, frames are gradually fixed to the structure. The movement and fixing of a frame in a specific position depends on its visual features, temporal information and the local neighbourhood of moving frames. Results of the ant-tree algorithm are clusters used in the decision process for classification of relevant/redundant segments. Common video segments which contain representative frames attached to the root of the tree are classified as relevant. The importance of relevant segments is defined by the number of frames from redundant segments in the corresponding cluster. Finally, a video summary is constructed by concatenating relevant video segments or representative key-frames. Additionally, the obtained summary is progressive, with richer descriptions deeper in the summary tree.

7. RESULTS AND CONCLUSIONS

In the early stage of the proposal, not having a complete system prototype yet, we could not perform any comprehensive test. However, general results for the modules allow us to draw preliminary conclusions with a good level of accuracy. For the face detection module, time performance is not an issue, since it has already been shown how the algorithm works in real-time: without the help from the tracking component, the face detection algorithm runs at 15 frames per second on a modest consumer PC [9], with high associated detection performance (around 92%-94%). The performance of the generic object detection system is difficult to evaluate, since in this scenario this task is rather subjective. Published results related to the proposed method validate the approach by classifying images containing only one out of four general categories (car, aeroplane, faces or motorbike), or background [7]. The results show an average classification rate of over 93% in this relatively simple task. The more challenging object localisation task, evaluated for the face category, achieves a poorer 60% accuracy, still giving useful hints on the relative position in the image.

Semantic segmentation has been tested the Microsoft Research Cambridge (MSRC) dataset of nine categories. The system can however be easily tailored on the specific scenario to be analysed, either indoor or outdoor. Indicative performance obtained with the base model on the MSRC dataset are presented in Table 1. Finally, we have tested our summarisation algorithm as a part of the collaborative system for automatic video summarisation in the TRECVID 2008 BBC rushes evaluation. Evaluation criteria included: IN – the fraction of inclusions found in the summary ($0 \div 1$); JU – “the summary contained lots of junk” (1 strongly agree – 5 strongly disagree), RE – “the summary contained lots of duplicates” (1 strongly agree – 5 strongly disagree). Our method scored: $IN = 0.4$, $JU = 3.4$ and $RE = 4$. The results show how the semantics that can be extracted from the video streams achieves an accuracy that enables an effective use in video editing scenarios. The system is highly expandable and audio processing data can be integrated, for example, for a richer scene semantic analysis.

8. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Seventh Framework Programme under grant agreement no. FP7-216444, PetaMedia.

9. REFERENCES

[1] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy – automatic naming of characters in tv video. In *BMVC*, 2006.

[2] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.

[3] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121 – 143, 2008.

[4] G. Passino, I. Patras, and E. Izquierdo. Image semantic labelling via heterogeneous low-level descriptors integration in a conditional random field. In *WIAMIS*, 2009.

[5] T. Piatrik and E. Izquierdo. Hierarchical summarisation of video using ant-tree strategy. In *CBMI*, 2009.

[6] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994.

[7] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[8] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006.

[9] D. Zhang, S. Z. Li, and D. Gatica-Perez. Real-time face detection using boosting in hierarchical feature spaces. In *ICPR*, 2004.

[10] S. Zhu and Y. Liu. Automatic scene detection for advanced story retrieval. *Expert Systems with Applications*, 36(3, Part 2):5976 – 5986, 2009.