Video Object Hyper-Links for Streaming Applications

Daniel Gatica-Perez[1], Zhi Zhou[1], Ming-Ting Sun[1], and Vincent Hsu[2]

[1] Department of Electrical Engineering, University of Washington
Seattle, WA 98195 USA
[2] CCL/ITRI
Taiwan

Abstract. In video streaming applications, people usually rely on the traditional VCR functionalities to reach segments of interest. However, in many situations, the focus of the people are particular objects. Video object (VO) hyper-linking, i.e., the creation of non-sequential links between video segments where an object of interest appears, constitutes a highly desirable browsing feature that extends the traditional video structure representation. In this paper we present an approach for VO hyper-linking generation based on video structuring, definition of objects of interest, and automatic object localization in the video structure. We also discussed its use in a video streaming platform to provide object-based VCR functionalities.

1 Introduction

Due to the vast amount of video contents, effective video browsing and retrieval tools are critical for the success of multimedia applications. In current video streaming applications, people usually rely on VCR functionalities (fast-forward, fast-backward, and random-access) to access segments of video of interest. However, in many situations, the ultimate level of desired access is the object. For browsing, people may like to jump to the next object of interest or fast-forward but only display those scenes involving the object of interest. For retrieval, users may like to find an object in a sequence, or to find a video sequence containing certain video objects. The development of such non-sequential, content-based access tools has a direct impact on digital libraries, amateur and professional content-generation, and media delivery applications [8].

VO hyper-linking constitutes a desirable feature that extends the traditional video structure representation, and some schemes for their generation have been recently proposed [5], [2], [13]. Such approaches follow a segmentation and region matching paradigm, based on (1) the extration of salient regions (in terms of color, motion or depth) from each scene depicted in a video shot, (2) the representation of such regions by a set of features, and (3) the search for correspondences among region features in all the shots that compose a video clip. In particular, the work in [2] generates hyper-links for moving objects, and the work in [13] does so for depth-layered regions in stereoscopic video. In [9], face

detection algorithms [15] were used to generate video hyper-links of faces. However, in spite of the current progress [12], automatic segmentation of arbitrary objects continues to be an open problem.

In this paper, we present an approach for VO hyper-linking generation, and discuss its application for video streaming with object-based VCR functionalities. After video structure creation, hyper-links are generated by object definition, and automatic object localization in the video structure. The object localization algorithm first extracts parametric and non-parametric color models of the object, and then searches in a configuration space for the instance that is the most similar to the object model, allowing for detection of non-rigid objects in presence of partial occlusion, and camera motion. As part of a video streaming platform, users can define objects, and then fast-forward, fast-reverse, or random-access based on the object defined.

The paper is organized as follows. Section 2 discusses the VO hyper-linking generation approach. Results are described in Section 3. Section 4 describes a streaming video platform with support for object-based VCR functionalities. Section 5 provides some concluding remarks.

2 VO hyper-link generation

2.1 Video structure generation

A summarized video structure or Table of Contents (TOC) (Fig. 1), consisting of representative frames extracted from video, cluster, shot, and subshot levels, is generated with the algorithms described in [6]. The TOC reduces the number of frames where the object of interest will be searched to a manageable number. Users can specify objects of interest to generate hyper-links, by drawing a bounding box on any representative frame.

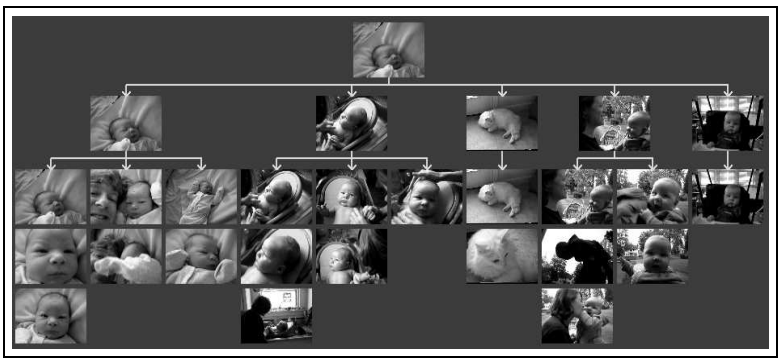

Fig. 1. Video Tree Structure. The root, intermediate, and column leaf nodes of the tree represent the video clip, the clusters, and the shots, respectively. Each image on a column leaf corresponds to frames extracted from each subshot.

2.2 Object localization as deterministic search

Object localization constitutes a fundamental problem in computer vision [15], [10], [18], [16], [3]. In pattern theory terms [7], [16], given a template (the image of an object) $\bar{I}(\mathbf{x})$ with support $\bar{\mathcal{D}} \subset \mathcal{R}^2$, any other image $I(\mathbf{x})$ that contains the object (with support $\mathcal{D} \subset \mathcal{R}^2$) can be considered as generated from the template $\bar{I}$ by a transformation T_X of the template into the image,

$$\bar{I}(\mathbf{x}) = I(T_X(\mathbf{x})), \mathbf{x} \in \bar{\mathcal{D}}, \tag{1}$$

where T_X is parameterized by X over a configuration space $\mathcal{X}$. In practice, Eq. 1 becomes only an approximation, due to modeling errors, noise, etc. In a deterministic formulation, localizing the template in a scene consists of finding the configuration $\hat{X} \in \mathcal{X}$ that minimizes a similarity measure $d(\cdot)$,

$$\hat{X} = \arg\min_{X \in \mathcal{X}} d_X = \arg\min_{X \in \mathcal{X}} d(I(T_X(\mathbf{x}), \bar{I}(\mathbf{x})). \tag{2}$$

We represent the outlines of objects by bounding boxes, and restrict the configuration space $\mathcal{X}$ to a quantized subspace of the planar affine transformation space, with three degrees of freedom that model translation and scaling. While far from representing complex object shapes and motions, the simplified $\mathcal{X}$ is useful to locate targets. The interior of an object could be approximately transformed by pixel interpolation using the scale parameter. Alternatively, one can define a similarity measure that depends not directly on the images, but on image representations that are both translation and scale invariant, so

$$\hat{X} = \arg\min_{X \in \mathcal{X}} d(f(I(T_X(\mathbf{x})), f(\bar{I}(\mathbf{x}))). \tag{3}$$

With this formulation, the issues to define are f, d, the search strategy, and a mechanism to declare when the objects is not present in the scene.

2.3 Reducing the search space with color likelihood ratios

Pixel-wise classification based on parametric models of object/background color distributions has been used for image segmentation [1] and tracking [14]. We use such representation to guide the search process. In the representative frames from which the object is to be searched, let $\mathbf{y}$ represent an observed color feature vector for a given pixel $\mathbf{x}$. Given a single foreground object, the distribution of $\mathbf{y}$ for such frame is a mixture

$$p(y|\Theta) = \sum_{i \in \{F,B\}} p(O_i) p(y|O_i, \theta_i), \tag{4}$$

where F and B stand for foreground and background, $p(O_i)$ is the prior probability of pixel $\mathbf{x}$ belonging to object O_i ($\sum_i p(O_i) = 1$), and $p(y|O_i, \theta_i)$ is the conditional pdf of observations given object O_i, parameterized by θ_i ($\Theta = \{\theta_i\}$). Each conditional pdf is in turn modeled with a Gaussian mixture [11],

$$p(y|O_i,\theta_i) = \sum_{j=1}^{M} p(w_j)p(y|w_j,\theta_{ij}), \tag{5}$$

where $p(w_j)$ denotes the prior probability of the j-th component, and the conditional $p(y|w_j,\theta_{ij}) = \mathcal{N}(\mu_{ij},\Sigma_{ij})$ is a multivariate Gaussian with full covariance matrix. In absence of prior knowledge $p(O_F) = p(O_B)$, and Bayesian decision theory establishes that each pixel can be optimally associated (in the MAP sense) to foreground or background by evaluating the likelihood ratio

$$\frac{p(y|O_F,\theta_F)}{p(y|O_B,\theta_B)} \underset{H_B}{\overset{H_F}{\gtrless}} 1 \tag{6}$$

The likelihood functions are on-line estimated using the Expectation- Maximization (EM) algorithm, the standard procedure for Maximum Likelihood parameter estimation [11]. Additionally, model selection is automatically estimated using the Minimum Description Length (MDL) principle.

RGB models are estimated when a new object is defined, and then applied to the set of representative frames in the video summary. An example is shown in Fig. 2. Only those pixels whose colors match the object color distribution are chosen as candidate search configurations. Finally, as the background color distribution is likely to change from shot to shot (possibly rendering low values for $p(y|O_B,\theta_B)$) probabilities are thresholded to ensure that candidate configurations truly correspond to object colors.

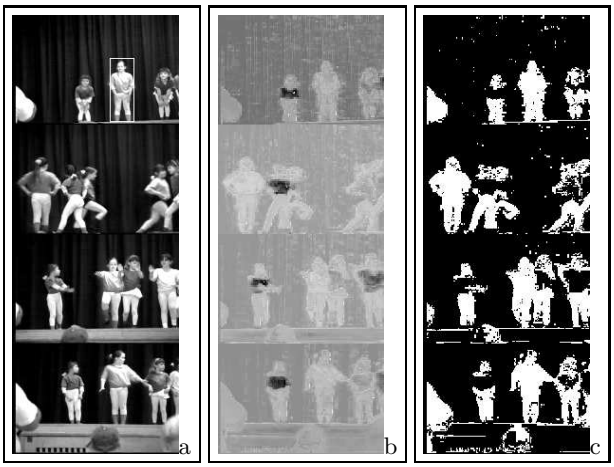

Fig. 2. Extraction of candidate configurations. *Dancing Girls* sequence. (a) Frames extracted from the video clips (the object has been defined by a bounding box). (b) Log-likelihood ratio image for learned foreground and background color models. Lighter gray tones indicate higher probability of a pixel to belong to the object. (c) Binarized image after decision. White regions will be used to generate candidate configurations.

2.4 Localization using Bhattacharyya coefficient

We use the color pdf of the interior of the configuration $X \in \mathcal{X}$ as the function $f(\cdot)$ in Eq. 3. Let $f(\bar{I})$ and $f(I_X)$ denote the color pdfs of the object and the configuration X, respectively. As discussed in [4], measuring similarity among two distributions can be defined as maximizing the Bayes error associated with them. The Bhattacharyya coefficient is a measure related to the Bayes error defined by

$$\rho_X = \rho(f(\bar{I}), f(I_X)) = \int (f(\bar{I}(\mathbf{x}))f(I_X(\mathbf{x})))^{1/2} d\mathbf{x} \tag{7}$$

and can be used to define a metric

$$d_X = (1 - \rho(\hat{f}(\bar{I}), \hat{f}(I_X)))^{1/2} \tag{8}$$

when the pdfs $f(\cdot)$ are represented by discrete densities $\hat{f}(\cdot)$. The discrete pdfs for model and candidate configuration are directly estimated by normalizing color histograms (3-D RGB, $8 \times 8 \times 8$ bins). Except for quantization effects, this color discrete density estimate is translation and scale invariant, unlike other representations, like color coocurrence histograms [3], which are translation invariant but not scale-invariant.

In the search, the translation component is quantized by a factor of 4 in each direction, and the scaling component is quantized to 5 different scales ranging between 0.5 and 2. If a whole QSIF image was to be searched, the number possible configurations would be 6600. We only search those positions with high likelihood as indicated in white regions in Fig. 2(c). Finally, the decision on the presence of the object is based on thresholding of the Bhattacharyya coefficient.

2.5 Video hyper-link generation

Hyper-links are constructed based on object detection/absence for each shot. If links are desired to the subshot level, the described object localization has to be applied on each of the leave frames in the TOC. Video browsing will occur by displaying the subshots for which the object was localized. Alternatively, hyper-links could be required only at higher levels of the hierarchy (shot, cluster). In that case, the object localization algorithm processes subshot frame leaves until it detects an object, and then jumps to the next shot or cluster, thus requiring less processing in average.

3 Results

Fig. 3 illustrates the results obtained in the *Girls* video, captured with a moving hand-held camera. One can observe that the algorithm has been able to detect the user-specified objects correctly, in presence of partial occlusion and change of size. Another detection example is shown in Fig. 4. We observe that detection of the object of interest has been correct, but also regions whose features can

not be discriminated are incorrectly labeled as object (false positives), and also the model might not discriminative enough, so multiple good matches occur. These results are obviously preliminary. Several issues are currently under study for object localization improvement, including the use of illumination-invariant object color models, the use of additional features, and the definition of a decision mechanism based on probability models of positive and negative examples.

Hyper-links are created, and the leaves in the TOC that contain the object are highlighted in the GUI, as shown in Fig. 5, allowing for fast browsing in the video structure besides the capability for video playing. The computational complexity is dependent on object size. In the current implementation without any optimization, it takes five seconds to search among 3000 configurations per QSIF image, on a Pentium III, 600 MHz PC. By off-line generation of the main objects in a video clip, the system can provide real time object-based browsing capabilities.

4 A streaming video system supporting Table of Contents and object-based VCR functionalities

A block diagram of a streaming video system is shown in Fig. 6. The system has a typical Server/Client structure. The video sequences are encoded in MPEG-4 and stored in the server with the associated metadata files. The system supports the conventional VCR functionalities such as Play, Pause, Random Access, Step Forward, Fast Forward, and Fast Reverse, plus the video TOC. The VCR functionalities are implemented as discussed in [10]. For simple implementation, we use I-pictures for random access, fast-forward, and fast-reverse. We are incorporating the object-based VCR functionalities into the system.

In the actual applications, the client connects to the remote server over an IP network, and selects the video stream of interest. Two types of logical channels are established between the server and the client: the control channel and the data channel. The TOC and the VCR commands are transmitted in the control channel, while the video packets are transmitted in the data channel. The server sends the TOC of the requested video sequence to the client. The TOC, containing clusters of the key frames of the sequence, is displayed as shown in Fig. 7. The client can choose to play from the beginning of the video sequence, or click on a frame in the the TOC to start playing from that particular segment. The VCR and Hyperlinking Manager receives the commands and retrieves the corresponding part of the video sequence, which is then sent by the Stream Manager to the client for decoding and displaying. The key frames in the TOC are mapped to the closest I-pictures to allow easy decoding. During the play of the video, the user can use the conventional VCR functionalities (e.g. fast-forward, fast reverse, etc.) to manipulate the play of the video. The user can also stop the video and jump to another key frame of interest in the TOC. With the incorporation of the object-based VCR functionalities, the user will be able to stop the video at any frame, define an object of interest in the frame, and use

Fig. 3. Object localization. *Girls* video sequence. (a), (b) and (c) illustrate the object localization process for three different user-defined video objects.

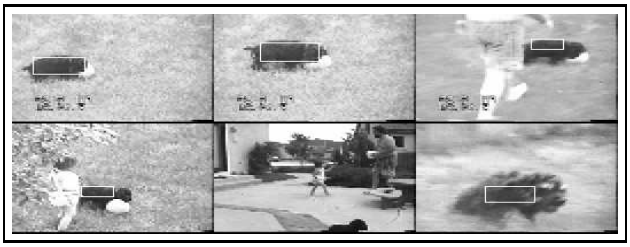

Fig. 4. Object Localization. *Dog* video sequence.

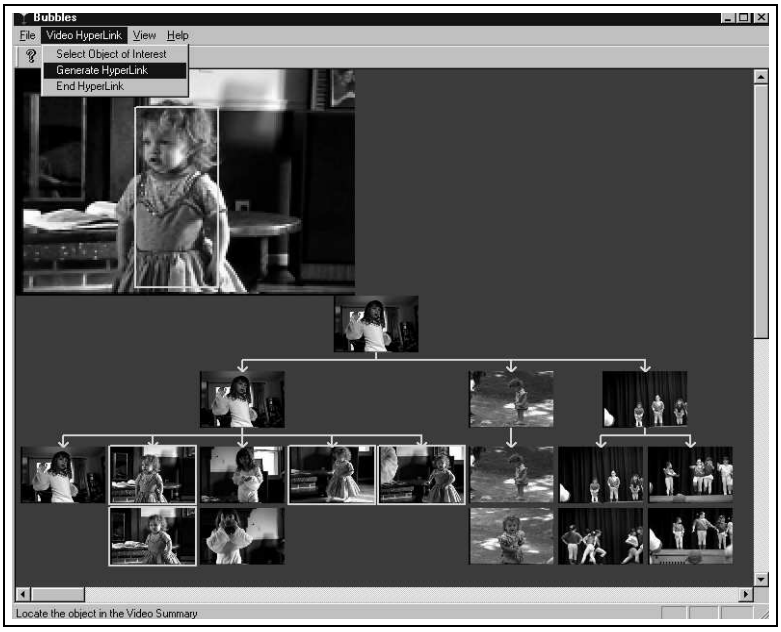

Fig. 5. VO hyper-link generation. The frames where the object has been detected are highlighted.

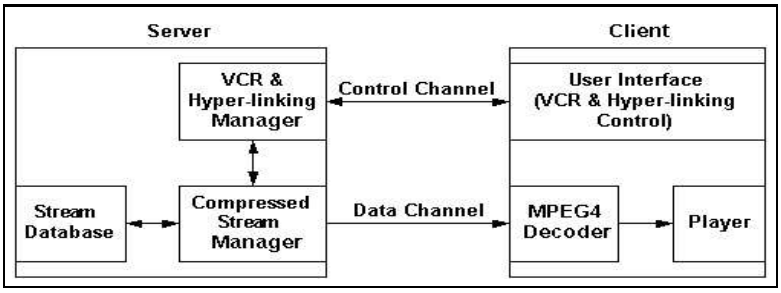

Fig. 6. Block diagram of streaming video system.

the object-based VCR functionalities through the support of the automatically generated VO hyper-links.

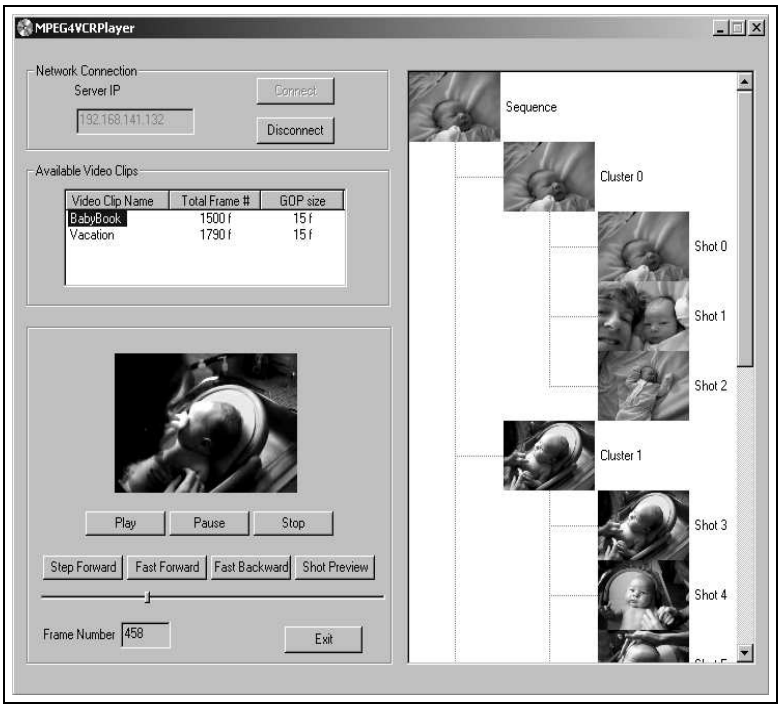

Fig. 7. Streaming video with VCR functionalities and Table Of Contents

5 Conclusions

We have presented a methodology to create video object hyper-links for object-based video streaming applications. Although the obtained results are encouraging, we acknowledge that object localization is a hard problem, and current efforts are directed to improve discriminitaion. We have implemented a streaming video system with Table of Content and VCR functionality support, and are incorporating the object-based VCR functionality features into the system.

Acknowledgements

The video sequences used in this study belong to the Eastman-Kodak Home Video Database©.

References

1. S. Belongie, C. Carson, H. Greenspan, and J. Malik, "Color and Texture Image Segmentation Using the Expectation-Maximization Algorithm and Its Application

to Content-Based Image Retrieval,", In *Proc. IEEE Int. Conf. Comp. Vis.*, Bombay, Jan. 1998.
2. P. Bouthemy, Y. Dufournaud, R. Fablet, R. Mohr, S. Peleg, and A. Zomet, "Video Hyper-links Creation for Content-Based Browsing and Navigation," in *Proc. Workshop on Content-Based Multimedia Indexing*, Toulouse, France, October 1999.
3. P. Chang and J. Krumm, "Object Recognition with Color Coocurrence Histograms," in *Proc. IEEE Int. Conf. on CVPR*, Fort Collins, CO, June 1998.
4. D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift," in *Proc. IEEE Conf. on Comp. Vis. and Patt. Rec.*, Hilton Head Island, S.C., June 2000.
5. Y. Deng, and B. S. Manjunath "Netra-V: Toward an Object-Based Video Representation," *IEEE Trans. on CSVT*, Vol. 8, No. 5, pp. 616-627, Sep. 1998.
6. D. Gatica-Perez, M.-T. Sun, and A. Loui, "Consumer Video Structuring by Probabilistic Merging of Video Segments," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Tokyo, Aug. 2001.
7. U. Grenander, *Lectures in Pattern Theory* Springer, 1976-1981.
8. C.W. Lin, J. Zhou, J.Youn, and M.T. Sun, "MPEG Video Streaming with VCR Functionality," *IEEE Trans. on CSVT*, Vol. 11, No. 3, pp. 415-425, Mar. 2001.
9. W.-Y. Ma and H.J. Zhang, "An Indexing and Browsing System for Home Video," In *Proc. EUSIPCO, European Conference on Signal Processing.* Patras, Greece, 2000, pp. 131-134.
10. J. MacCormick and A. Blake, "A probabilistic contour discriminant for object localisation," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 390-395, 1998.
11. G.J. MacLachlan and D. Peel. *Finite Mixture Models.* John Wiley and Sons, N.Y., 2000.
12. M. Meila and J. Shi, "A random walks view of spectral segmentation," in *Proc. Eighth Int. Workshop on AI and Stats,* Jan. 2001.
13. K. Ntalianis, A. Doulamis, N. Doulamis, and S. Kollias, "Non-Sequential Video Structuring Based on Video Object Linking: An Efficient Tool for Video Browsing and Indexing," in *Proc. IEEE Int. Conf. Image Processing*, Thessaloniki, Greece, October 2001.
14. Y. Raja, S. McKenna, and S. Gong, "Colour Model Selection and Adaptation in Dynamic Scenes," in *Proc. ECCV*, 1998.
15. H. Rowley, S. Baluja, and T. Kanade, "Human Face Detection in Visual Scenes," Tech. report CMU-CS-95-158R, Computer Science Department, Carnegie Mellon University, November, 1995.
16. J. Sullivan, A. Blake, M. Isard and J. MacCormick, "Object Localization by Bayesian Correlation," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 1068-1075, 1999.
17. H.J. Zhang, "Content-based Video Browsing and Retrieval," In B. Fuhrt, Ed., *Handbook of Multimedia Computing*, CRC Press, Boca Raton, 1999, pp. 255-280.
18. Y. Zhong and A. K. Jain, "Object Localization Using Color, Texture and Shape," in *Proc. Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition,* Venice, pp. 279-294, May 1997: