IDIAP RESEARCH REPORT

Linking Objects in Videos by Importance Sampling

Daniel Gatica-Perez [a] Ming-Ting Sun [b]

IDIAP–RR 02-20

May 6, 2002

to appear in
IEEE International Conference on Multimedia and Expo 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, Martigny, Switzerland
[b] Dept. of Electrical Engineering, University of Washington, Seattle, WA, USA

IDIAP Research Report 02-20

Linking Objects in Videos by Importance Sampling

Daniel Gatica-Perez Ming-Ting Sun

May 6, 2002

to appear in
IEEE International Conference on Multimedia and Expo 2002

Abstract. We present an approach to create hyper-links between video segments that contain objects of interest, based on video structuring, object definition, and stochastic object localization in the video structure. Localization is formulated in the Metric Mixture model framework, which allows for the joint probabilistic modeling of a (user-defined) set of color appearance exemplars and their geometric transformations. Candidate object configurations are drawn from a prior distribution using importance sampling -which guides the search towards regions of the configuration space likely to contain the correct object configuration, thus avoiding exhaustive processing- and evaluated using Bayes' rule. Results of linking real objects (with changes of size and pose) in several home videos illustrate the performance of the method.

1 Introduction

The development of non-sequential tools for content-based video browsing and retrieval has a direct impact in digital libraries, amateur and professional content generation, and media delivery applications. The first step in this direction has been the automatic generation of video structure, which allows for browsing functions at different levels (shots, clusters), but limited to the image level. However, the ultimate level of desired access is the object. In this view, hyper-links between video segments that contain objects of interest constitute a valuable feature [1], [11] that effectively complements the video structure representation. Indeed, as the number of frames in a video summary increases, users become less motivated to interact with it. The capability of jumping backwards and forward in time to browse video based on user-defined objects provides more focused interaction.

Schemes for video object hyper-linking have been recently proposed [1], [9], [11]. Moving object hyper-links are generated in [1]. The work in [11] does so for depth-layered regions in stereoscopic video. In [9], face detection algorithms were implemented [12] to generate face hyper-links. However, real objects are not motion-consistent, and object segmentation continues to be an unsolved problem, in spite of progress.

In this paper, we propose an approach to create video object hyper-links based on three steps: video structuring, object definition, and stochastic object localization in the video structure. Localizing objects is a fundamental problem in computer vision [14], [12], [10], [2]. In brief, given a discriminative object representation, localization is a search problem in a configuration space, clearly demanding if the latter is large or continuous [12], [2]. One distinctive feature of object localization for hyper-linking is the fact that -in order to make it truly interactive- objects should be allowed to be defined on-the-fly, which imposes constraints on learning and inference schemes. We formulate the solution in the recently proposed *Metric Mixture* model [15], which allows for the joint probabilistic modeling of *exemplars* and their geometric transformations in a space that has no vector structure. The probabilistic formulation is appealing as uncertainty is dealt with in a principled basis. Exemplars are object representations that can be readily extracted from raw data; in our case, they correspond to color image templates that define an object of interest. After defining the configuration space of our problem, we address object localization by random sampling from the object prior distribution, [10], [13]. Candidate configurations are drawn using *importance sampling* , [6], [8] -which guides the stochastic search process towards regions of the configuration space likely to contain the true object configuration, thus avoiding exhaustive processing- and evaluated using Bayes' rule. To this purpose, we define an *importance function* based on parametric and non-parametric object color models. We illustrate the performance of our approach with real video objects that have variations of pose and size, extracted from a home video database.

The paper is organized as follows. Section 2 describes the video structuring step. Section 3 presents the object localization algorithm. Section 4 describes the hyper-link generation algorithm. Section 5 presents results. Section 6 provides some concluding remarks.

2 Video Structure Generation

A summarized video structure (Fig. 3), consisting of representative frames extracted from video, cluster, shot, and subshot levels, is generated as described in [5]. While the number of frames depends on shot appearance variation, the summary maintains a manageable number for object localization. A hidden, additional set of frames is available for further search if necessary. Users specify objects of interest directly on the representative frames or while playing the video, by drawing a bounding box around it (Fig. 1(a)). The process can be repeated in other frames where the object appears, to create a small set of color image templates, called exemplars in the following.

3 Object Localization in Metric Spaces

In pattern theory terms [7], an observed image $z \in \mathcal{Z}$ can be approximated as a template $x \in \mathcal{X}$ on which a continuous geometric transformation $t \in \mathcal{T}$ has been applied, $z \approx tx$. If the discrete set $\mathcal{X}$ represents an object model, $\mathcal{X} \times \mathcal{T}$ describes the object and its possible transformations. The representation is attractive: while $\mathcal{T}$ can model global transformations, $\mathcal{X}$ can represent complex variations of shape, appearance, pose, etc.

A probabilistic formalization of this approach was developed in [4], and generalized in [15] for non-vector exemplars. Exemplars are convenient low-level object representations (color or edge image templates) because they can be extracted relatively easily from images, and then used to define object models, without resorting to complex intermediate representations. However, several useful operators to compare exemplars do not correspond to operations in a vector space. For instance, the histogram intersection [14] between two image templates does not constitute a norm in $\mathbb{Z}^2$. For object tracking purposes, the work in [15] described in a principled way how probabilistic mixture models can be defined and learned from exemplars in a metric space.

We propose to use a similar formulation for object localization. In our case, $\mathcal{X}$ is defined by the set of user-defined color image templates $\tilde{x}_k$ that model object appearance, $\mathcal{X} = \{\tilde{x}_k, k = 1, ..., K\}$, equipped with a distance function ρ. Additionally, the transformation space $\mathcal{T}$ is defined as a subspace of the euclidean transformations that models translation and scaling, which is useful to locate targets. Elements of the exemplar-transformation space will be denoted by the pair $X = (k, t)$.

3.1 Formulation of the localization problem

In similar fashion to [10], [13], we formulate object localization using Bayesian theory and stochastic simulation. Given a prior distribution on the possible object configurations, denoted by $p(X)$, an observed image z, and an observation likelihood $p(z|X)$, the posterior can be expressed by Bayes' rule as

$$p(X|z) \propto p(z|X)p(X). \tag{1}$$

There are well-known Monte Carlo discrete representations for distributions, discussed elsewhere [8], [10]. In brief, a posterior can be approximated by a set of weighted samples (also called particles) $\{(X^{(i)}, \pi^{(i)}), i = 1, ..., N\}$. From this approximation, inference about X can be done. In a vector space [10], [13], moments of the posterior can be readily computed. In contrast, averages are not defined in a space that has no vector structure. However, peaks in the posterior still provide evidence of object location. Therefore, our approach for localization draws a set of random proposals S from the prior, evaluates the observation likelihood at each proposal (extracting image measurements, and in fact quantifying discrepancy between the prior from which candidates were sampled and the true posterior), and displays the configuration that maximizes the posterior in the sample set,

$$X^* = \arg \max_{X^{(i)} \in S} p(z|X)p(X) \tag{2}$$

With this formulation, the distributions and a decision rule to decide whether the object is present have to be specified.

3.2 Modeling the observation likelihood

The observation likelihood is modeled by a *metric exponential* distribution [15]

$$p(z|X) = p(z|k, t) \propto \frac{1}{\mathbf{Z}} e^{-\lambda \rho(z, t\tilde{x}_k)}, \tag{3}$$

where $t\tilde{x}_k$ corresponds to a transformed exemplar, $\mathbf{Z}$ is a normalization constant, and λ is a parameter that has to be estimated from data. It is easy to show that, assuming that t and k are

independent, the conditional likelihood on transformations $p(z|t) = \sum_K p(z,k|t) = \sum_K p(k)p(z|k,t)$, i.e., it is a mixture of metric exponentials, whose centers are the transformed exemplars $t\mathbb{A}_k$, and whose weights are given by the exemplar prior $p(k)$ [15]. We further assume a quadratic form as a reasonable noise model, when ρ is the distance function based on the Bhattacharyya coefficient[1] [3]. In that case, the exponential parameter and the normalization constant can be approximated by $\lambda \approx \frac{1}{2\sigma^2}$ and $\mathbf{Z} \propto \sigma^d$, where σ is a "variance" parameter [2] that measures the spread of the metric exponential "around" its center, and d is a measure of the "effective" dimensionality of the unknown exemplar space.

The chosen distance, denoted by ρ_{BT}, is defined by

$$\rho_{BT}(z, t\mathbb{A}_k) = (1 - d_{BT}(f(z), f(t\mathbb{A}_k)))^{1/2}, \tag{4}$$

where $f(t\mathbb{A}_k)$ denotes a 4-D normalized histogram (color + relative position) of the transformed exemplar, and the Bhattacharyya coefficient is defined by $d_{BT} = \sum (f(z)f(t\mathbb{A}_k))^{1/2}$. Except for quantization effects, the normalized histogram is translation- and scale- invariant, unlike other representations, like coocurrence histograms [2], which are not scale-invariant.

For tracking purposes, exemplars are clustered to reduce the model complexity of the observation likelihood, and parameter estimation is performed from hundreds of examples [15]. However, for video object hyper-linking, users usually specify one or a handful of exemplars, so clustering and estimation from such small amount of data are not possible. Instead, we have estimated the parameters for several objects of interest on training videos, and used the same parameters for all new cases (see Section 5). Additionally, each of the user-specified exemplars is treated as a center in the Metric Mixture model. More satisfactory solutions are currently under study.

3.3 Importance sampling from the prior

The prior distribution $p(X)$ encodes the knowledge about object location. As stated before, exemplar indices and geometric transformations are independent, so $p(X) = p(k,t) = p(k)p(t)$. The most general assumption is a uniform distribution on both exemplar index and geometric transformations (in the latter case, over a finite interval). However, knowledge about possible locations at each representative frame can be extracted using object features, like color or texture, and could be useful to guide the random search. This is properly modeled through the use of importance sampling [6], [8]. This is a technique aimed at improving the efficiency of simulation methods, and useful when such additional knowledge can be expressed by a (normalized) importance function $g(X)$ that emphasizes the regions of the configuration space which contain more information about $p(X)$. The technique first draws random samples $X^{(i)}$ from $g(X)$ rather than from $p(X)$, which concentrates particles in better proposal regions, and then introduces a correction mechanism in order to keep the particle set as a faithful representation of $p(X)$. Such correction takes the form of an importance ratio factor defined by $p(X = X^{(i)})/g(X = X^{(i)})$, and applied in the particle weights $\pi^{(i)}$. The introduction of the importance ratio guarantees that sampling from $g(X)$ has (asymptotically) null effect on the consistency of the discrete representation of $p(X)$ [8]. In our work, we keep the assumption of uniformity on the exemplar index distribution, $p(k) = u(k)$, and use importance sampling to draw samples from the geometric transformation distribution $p(t)$.

3.4 Constructing the importance function

We use a parameteric color model of each exemplar to generate candidate configurations in each of the frames on which the object is searched for. Let y represent an observed color vector for a given pixel. Given a single foreground object, the distribution of y is a mixture, $p(y|\Theta) = \sum_{i \in \{F,B\}} p(O_i)p(y|O_i, \theta_i)$,

[1] ρ_{BT} is a true metric in functional space, but not in exemplar space.
[2] a true variance for certain distance functions.

where F and B stand for foreground and background, $p(O_i)$ is the prior probability of pixel of color y of belonging to object O_i, and $p(y|O_i, \theta_i)$ is the conditional pdf of observations given object O_i, represented by a Gaussian mixture model (GMM) [6]. In absence of prior knowledge ($p(O_F) = p(O_B)$), the optimal classification into foreground/background is obtained by comparing the likelihood ratio $\frac{p(y|O_F,\theta_F)}{p(y|O_B,\theta_B)}$ to 1.

Color models are on-line estimated for each exemplar using the Expectation-Maximization (EM) algorithm [6]. Then, for each searched image, a binary image I_k^b is built based on pixel classification, followed by morphological processing, in order to generate blobs whose colors match the object model. 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 the generated blobs truly correspond to object colors. Finally, a blob image is obtained by computing the maximum of the binary images obtained for each exemplar, $I^g = \vee_{k=1}^{K} I_k^b$. An example is shown in Fig. 1 (a) and (b), for one exemplar.

Recall that the transformation space $\mathcal{T}$ has been chosen as a subset of the euclidean transformations, allowing for translation (o) and scaling (s), so any $t \in \mathcal{T}$ can be denoted by $t = (o, s)$. Assuming independence and a uniform distribution for the scaling parameter, the importance function is $g(t) = g(o, s) = g(o)u(s)$. To specify a functional form for the translation parameters $g(o)$, we use the binary image I^g. We define $g(o)$ as a GMM,

$$g(o) = g(o|\Phi) = \sum_{i=1}^{C} p(c_i)p(o|c_i, \phi_i), \tag{5}$$

where c_i denotes each of the C connected components of I^g. The parameters ϕ_i correspond to the mean and 2-D covariance matrix of the pixels in each component. Furthermore, the prior distribution $p(c_i)$, which defines the relative contribution of each blob to the mixture, is determined from two features: the blob size, and its maximum color similarity (i.e. the minimum distance ρ_{BT}) to the exemplars that define the object model,

$$p(c_i) = \sum_{j \in \{size, color\}} p(w_j)p(c_i|w_j), \tag{6}$$

The distributions $p(c_i|w_{size})$ and $p(c_i|w_{color})$ are estimated directly from data. Finally, the prior $p(w_j)$ is assumed uniform. Random sampling will draw more configurations from large blobs whose color distribution match better the object's (Fig. 1(c)).

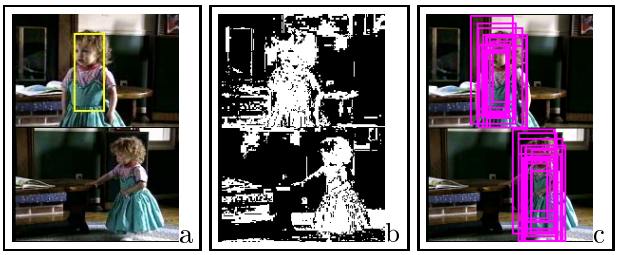

Figure 1: (a) Frames from *Girl* video (the template defined in the first frame constitutes the exemplar X_1).(b) Binary image after pixel classification. (c) Importance sampling. The displayed samples are the ones with largest posterior (Eq. 2). Only the transformed template contours are shown.

3.5 Object detection/absence

The described method outputs both the geometric transformation and the exemplar that best match the object model for each representative frame. Object absence is decided based on thresholding of $p(z|X)p(X)$. The threshold is empirically determined from a set of positive and negative examples [12].

4 Video Hyper-Link Generation

Hyper-links are constructed based on object detection/absence for each subshot. The leaves in the video structure that contain the object are highlighted, as shown in Fig. 3, and the subshots in which the object was localized are displayed. Alternatively, hyper-links could be required only at higher levels of the hierarchy (shot, cluster). In that case, the algorithm is applied until it detects an object, and then moves to the next shot or cluster, requiring less processing in average.

5 Results

The results presented in this section correspond to the one-exemplar case ($K = 1$). Performance with multiple exemplars will be reported elsewhere. The RGB space is used for computation of normalized histograms ($8 \times 8 \times 8$ bins), and parametric models. For histograms, an additional dimension that indicates relative position (3 bins) is used to model spatial structure.

Table 1 shows the estimated parameters for the metric exponential distribution, for three video objects extracted from a home video database [5], and for all the distance measurements combined. Forty hand-labeled exemplars were used for each video object. The variation in the parameters is significant depending on the content, as objects with less variation throughout a video tend to generate sharper observation likelihoods.

Sequence	d	σ	λ
Wedding (Man)	16.60	0.20	11.91
Wedding (Bride)	9.05	0.21	11.57
Girl	24.06	0.15	22.22
All video objects	14.34	0.16	18.17

Table 1: Estimated parameters. Metric exponential model.

Fig. 2(b-e) illustrates the video object localization results obtained in several clips captured with a moving hand-held camera, when drawing 300 random samples from the prior. The range for the scale parameter was [0.5, 2]. The methodology has correctly detected the specified objects, in presence of partial occlusion and change of size and pose. Similar results have been obtained on other video objects and sequences in our database. As a byproduct, the method could be used to initialize an stochastic tracker [8], [10] for further video analysis. For comparison, Fig. 2(a) shows the best ten results obtained with exhaustive search, with translation quantized by a factor of 4 in each direction, and scaling quantized to 10 levels (13200 configurations). The computational complexity is dependent on object size. After color model estimation, it takes approx. 0.5 s. to process 300 samples per QSIF image, on a Pentium III, 600 MHz PC. This figure could be significantly improved by multi-resolution processing, and code optimization.

We are currently testing the performance of the method in a larger database (both in terms of number of video objects and total number of frames) using recall-precision measurements. Most object localization methods have been designed for specific object classes [12], or tested in controlled image databases [2]. We are not aware of any study of performance evaluation of arbitrary object localization algorithms in realistic environments.

The obtained results are encouraging. However, object localization based on one or a few examples in unrestricted scenes is a very hard problem. Our approach is limited by two factors: the discrimination that can be obtained with color histograms (as can be seen in the *Man* sequence), and the quality of the importance function. Several issues are currently under study, including the use of illumination-invariant color models and additional features, and the definition of a decision mechanism based on probability models of positive and negative detections.

Figure 2: Object localization by (a) exhaustive search, and (b-g) importance sampling (only the best configuration is shown). (b) *Girl*. (c) *Bride*. (d) *Man*. (e) *Maid of Honor*, (f) *Dancer*, (g) *Boy*. A single exemplar is specified in the first frame.

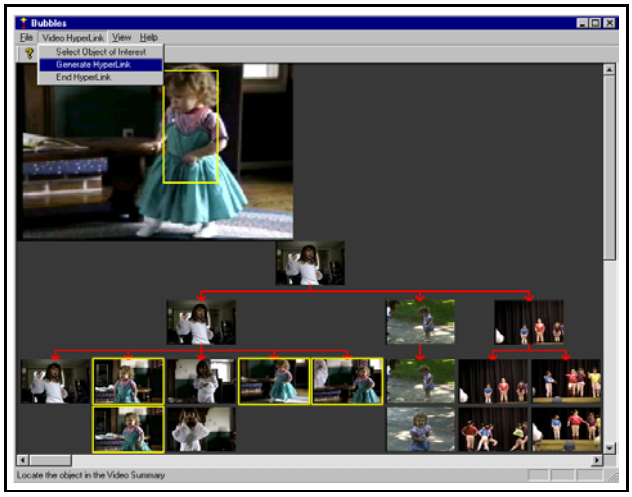

Figure 3: Video tree structure (sequence, cluster, shot and subshot nodes) and object hyper-links (highlighted images).

6 Concluding Remarks

We have presented a principled methodology to create video object hyper-links based on a probabilistic formulation of object localization, a process of random search in a product configuration space of exemplars and geometric transformations that considerably reduces computational complexity, while keeping good localization features. Results of good quality for object-based video browsing in real home videos have been obtained.

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

References

[1] 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 CBMI*, Toulouse, October 1999.

[2] P. Chang and J. Krumm, "Object Recognition with Color Coocurrence Histograms," in *Proc. IEEE CVPR*, Fort Collins, CO, June 1999.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift," in *Proc. IEEE CVPR.*, Hilton Head Island, S.C., June 2000.

[4] B. Frey and N. Jojic, "Learning graphical models of images, videos and their spatial transformations," in *Proc. UAI*, 2000.

[5] D. Gatica-Perez, M.-T. Sun, and A. Loui, "Consumer Video Structuring by Probabilistic Merging of Video Segments," in *Proc. IEEE ICME*, Tokyo, Aug. 2001.

[6] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, *Bayesian Data Analysis.* Chapman and Hall, 1995.

[7] U. Grenander, *Lectures in Pattern Theory.* Springer, 1981.

[8] M. Isard, and A. Blake, "ICondensation: unifying low-level and high-level tracking in a stochastic framework," in *Proc. ECCV*, 1998.

[9] W.Y. Ma and H.J. Zhang, "An Indexing and Browsing System for Home Video," in *Proc. EUSIPCO.* Patras, 2000.

[10] J. MacCormick and A. Blake, "A probabilistic contour discriminant for object localisation," in *Proc. IEEE ICCV*, pp. 390-395, Bombay, Jan. 1998.

[11] K. Ntalianis, A. Doulamis, N. Doulamis, and S. Kollias, "Non-Sequential Video Structuring Based on Video Object Linking," in *Proc. IEEE ICIP*, Thessaloniki, October 2001.

[12] H. Rowley, S. Baluja, and T. Kanade, "Human Face Detection in Visual Scenes," TR-CMU-CS-95-158R, Nov. 1995.

[13] 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.

[14] M.J. Swain and D. Ballard, "Color Indexing," *Int. J. of Comp. Vis.*, Vol. 7, pp. 11-32, 1991.

[15] K. Toyama and A. Blake, "Probabilistic Tracking in a Metric Space," in *Proc. IEEE ICCV*, Vancouver, Jul. 2001.