Motion Based Retrieval of Dynamic Objects in Videos

Che-Bin Liu
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
cbliu@uiuc.edu

Narendra Ahuja
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
ahuja@vision.ai.uiuc.edu

ABSTRACT

Most existing video retrieval systems use low-level visual features such as color histogram, shape, texture, or motion. In this paper, we explore the use of higher-level motion representation for video retrieval of dynamic objects. We use three motion representations, which together can retrieve a large variety of motion patterns. Our approach works on top of a tracking unit and assumes that each dynamic object has been tracked and circumscribed in a minimal bounding box in each video frame. We represent the motion attributes of each object in terms of changes in the image context of its circumscribing box. The changes are described via motion templates [4], self-similarity plots [3], and image dynamics [9]. Initially, defined criteria of the retrieval process are interactively refined using relevance feedback from the user. Experimental results demonstrate the use of the proposed motion models in retrieving objects undergoing complex motion.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*motion*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, relevance feedback*

General Terms

Algorithms, Experimentation

Keywords

Content-Based Video Retrieval, Motion Analysis

1. INTRODUCTION

Recently, some motion representations have been proposed to recognize different motion patterns such as human gaits, activities, periodic motions and texture motions. However, the existing content-based video retrieval (CBVR) approaches focus on low-level motion features such as pixel-level optical flow or affine parameters for motion content indexing, in addition to other visual features such as color, shape or texture. The main disadvantage of using low-level motion features lies in the recognition of complex motion patterns such as gaits. Such complex motion patterns can be effectively tackled using higher-level motion representations, which might be region based or image based, for example. But such an extension is not straightforward for video retrieval and often depends on many assumptions.

1.1 Motivation and Approach

According to the motion classification tree of objects proposed by Kambhamettu et al. [5], most real-world motions can be classified as rigid, articulated, elastic (deformable motion with topological invariance), or fluid. For example, vehicle movement is a rigid motion; animal/human movements are articulated motions in general; deformable objects affected by external force such as a dropping sheet of paper exhibit elastic motion; motions exhibiting topological variations and turbulent deformations are viewed as fluid motion.

These different types of movements become apparent via different characteristics of motion extracted from images and can then be used for retrieval. For instance, articulated motion can be characterized through periodicity of motion observed in videos and can be found in many biological movements, such as human or animal gaits. Apart from gaits, articulated motion also includes interesting kinds of movements, such as moving body parts like a man swimming in sports video, that people are interested in querying. These movements are generally localized in nature and can be characterized through region based motion features like motion presence and motion recency, which we will discuss in the later sections of the paper. Elastic and fluid motion, on the other hand, varies continuously across objects. The difference between these two classes of motions lies in the continuity of the object itself. When observing these types of motions, people usually have prior knowledge about the object and pay attention to the deformations or topological changes of the observed object. Either deformation or topological change is usually an important signature of object identity. Rigid motion characterizes poses and translations of rigid objects and it corresponds to affine parameter estimation in image analysis that has no information about object identification. To recognize a rigid object, people consult to shape, color or other visual cues other than motion to determine the object class.

MM'04, October 10-16, 2004, New York, New York, USA.

Based on the aforementioned visual properties, we investigate four motion properties: periodicity, presence, recency, and image dynamics. To represent these movements, we also adopt appearance based methods, rather than model based approaches, because (1) the types of dynamic objects of interest are unknown, and (2) the dynamic objects present in the video database may have huge variety so that no single model fits well all dynamic objects. In this paper, we represent motion periodicity using modified similarity plots [3] where we use normalized cross-correlation to measure image similarity. Then we proceed to represent motion presence and motion recency using temporal template approach [4] that reveals the tendency of movement and have been successfully used to recognize human activities. Following which we use image-level dynamics, motivated by [9] that characterizes the temporal changes between image frames, to capture variations of the object motion without object models. These motion properties can be harnessed to cover a wide range of interesting motion patterns and can be used to retrieve videos by queries that analyze the high-level motion content in videos.

Our system makes two assumptions. First, we assume that the dynamic objects in an image sequence have been tracked so that a minimal bounding box circumscribed for each dynamic object is available in any given video frame. Second, we assume that there is only one foreground object in each circumscribing box, and that the backgrounds do not change significantly over a short time period. There have been some tracking techniques which find a minimal bounding box for a moving object. Although in some cases tracking methods might fail to locate moving objects, we maintain the first assumption by interactively working with tracking methods. The second assumption can be removed if the dynamic object can be automatically segmented. However, this is a very difficult problem, especially when the types of moving objects are unknown. Therefore, our second assumption ensures that the dynamic objects in corresponding bounding boxes can be reasonably matched (or aligned). With these two assumptions, our objective becomes: given a sequence of bounding boxes whose changes in the image context representing the motion properties of a dynamic object, we find similar dynamic object sequences based on the similarity of changes in their image contexts.

2. METHOD

Assume that a dynamic object has been tracked and a minimal bounding box around the object in any give video frame is available. Note that the bounding boxes of a moving object may have different sizes in different video frames. Therefore, we first align the bounding boxes so that the appearances of the object are best matched. We then capture the motion content of the bounding box sequence using three representations described later. To retrieve similar sequences, we first compute the similarity measures between the representations of the given sequence and those in database. The similarity between sequences is interactively refined by integrating similarity measures according to user's feedbacks.

2.1 Image Alignment of Dynamic Objects

Normalized cross-correlation (NCC) [6] is used to match the appearances of dynamic objects in the bounding boxes. This method shifts a template image over a search image, measuring normalized cross-correlation at each point. The point associated with the maximal NCC value is selected as the best match. Although the original method requires that the search image be larger than the template image in both dimension, we extend the search region to include some neighbor pixels of the target bounding box. To help finding best match by NCC, we place spatial constraints on the search window to prune the candidate matches. This alignment process is fully automatic. The computations for all motion representations in the following are done in overlap regions of bounding boxes.

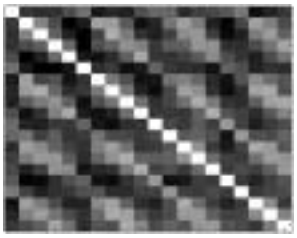

Figure 1: The similarity plot of a human running sequence.

2.2 Self-Similarity Plot

Cutler et al. [3] developed an approach to detection of periodic motion by using similarity plots, where they analyze periodic signals using an auto-correlation function. The idea is to use similarity plot to encode the projection of spatio-temporal dynamics of moving objects, and then analyze similarity plot for object classification.

Figure 1 shows the similarity plot of a twenty-frame human running sequence. The value at pixel (x, y) of the plot represents the similarity, defined in this paper as the value of normalized cross-correlation, between overlap regions of the bounding boxes in image frame x and y. The bright diagonals in the plot indicate periodic motion in the given sequence.

Two features of a similarity plot are defined in this paper. These are concerned with if the given sequence is periodic, and second, the length of a periodic cycle. For instance, as indicated in Figure 1, the motion of the object is periodic with a cycle length about six image frames, which corresponds to a half gait, or a stride.

2.3 Temporal Templates

Davis et al. [1, 4] introduced two temporal templates, motion-energy image (MEI) and motion-history image (MHI), to respectively represent the presence and the recency of object movement. Let $D(x, y, t)$ be a binary value indicating regions of motion at frame t. An MHI H_τ is defined as

$$H_\tau(x,y,t) = \begin{cases} \tau, & \text{if } D(x,y,t) = 1; \\ max(0, H_\tau(x,y,t-1) - 1), & \text{otherwise,} \end{cases}$$

where τ denotes the desired length of history. An MEI E_τ is defined as

$$E_\tau(x,y,t) = \begin{cases} 0, & \text{if } H_\tau(x,y,t) = 0; \\ 1, & \text{if } H_\tau(x,y,t) > 0. \end{cases}$$

In our system, we obtain D using image differencing. To distinguish between motion patterns, seven of Hu's moments are computed over MHIs and MEIs respectively, which are

translation- and scale-invariant. Then the Mahalanobis distance of Hu's moments of two MHIs or MEIs is used to measure the similarity between motions of τ image frames. Note that different motion patterns may need different lengths of history to be best described by these temporal templates. Therefore, we compute MHI and MEI with four different lengths ($\tau = 5, 10, 15, 20$).

Image sequence registration is a problem when using this method to compare motion patterns in two sequences. We overcome this problem by shifting one sequence and computing motion similarity for every possible sequence alignment. The measurement of the best similarity in temporal templates for the eventual similarity integration is chosen.

2.4 Image Dynamics

We model image-level dynamics in image subspace, which is similar to Soatto's [9] and Brand's [2] approaches. The image subspace is spanned by a set of basis images. The input image sequence is projected onto the subspace frame by frame and the projections form a trajectory in the subspace. We model the evolution of this trajectory using a first-order auto-regressive (AR) model. Therefore, the temporal behavior of an image sequence is captured by the evolution of the moving trajectory in image subspace.

Assume that we have n frames in an image sequence, and each image frame of the sequence is represented as a column vector $I_i \in \mathbf{R}^m$ in the raster scan order. Let μ be the mean of the images and $I'_i = I_i - \mu$. We use a matrix $X = [I'_1 I'_2 \dots I'_n]$ to denote the whole input image sequence around the mean image. Using the algorithm in [10], we find the eigenvectors $\{e_j\}_{j=1\dots k}$, which correspond to the largest k eigenvalues, of the covariance matrix XX^T. Therefore, we represent each image frame as $I_i = VP_i + \mu$, where $V = [e_1 e_2 \dots e_k]$ and $P_i = V^T(I_i - \mu)$. P_i is the projection of I_i in the subspace spanned by V. Furthermore, we treat the projections P_i as the k-dimensional random vectors observed at equal time intervals. The first-order k-variate AR model is defined as $P_i = AP_{i-1} + n_i$. The matrices $A \in \mathbf{R}^{k\times k}$ are the coefficient matrices of the AR model, and the k-dimensional vectors n_i are uncorrelated random noise with zero mean. Note that the AR model for each sequence are defined in different subspaces. Therefore, to measure the similarity between image sequences, we compute Martin's distance between AR models defined by $\{A, V\}$ pairs [8].

2.5 Integration of Similarity Measures

The respective similarity measures for three motion representations are all integrated to measure motion-content similarity in the circumscribing boxes of dynamic objects. Since the quality of retrieval results is subjective to user's visual perception, the ways to integrate different similarity measures may vary depending on the dynamic object of query. There have been some systems that require users to specify weights for their queries, which often leads to unsatisfactory results. In our system, we linearly combine similarity measures and dynamically adjust their weights according to user's interactive feedback [7]. Such relevance feedback based retrieval approach has been empirically proved to be very effective.

3. EXPERIMENTAL RESULTS

The video clips used in our experiments were randomly collected from TV programs or recordings of street scenes. In most cases, the videos involve camera motions. Currently, fifty image sequences have been used in our experiments. All images are converted into gray-level before we apply the algorithms. The retrieval results of five queries of different motion patterns are shown in figure 2.

The first test is a human walking video. Such motion pattern is periodic and has important signatures in motion presence and motion recency, where the MEI shows the motion is around the lower part of the object and the MHI indicates the major movement is toward right. The third best sequence of this query involves walking with an angle to the camera plane, which decreases the similarity in motion presence to the test sequence. The forth best sequence involves walking toward the opposite direction of the test sequence, and the features for motion recency show the differences.

In the second test, we use a tennis video where a player back to the camera performs a right-handed swing including his follow-through. This is a full body motion, but the motion in image context has emphasis on arm swing (including the racket) and leg movement. The similarity measures in motion presence and motion recency are most relevant in retrieving similar videos. Note that the videos of left-handed swings or swings of a player facing the camera are not retrieved.

The third test video is about the movement of a bird's wings. Although the motion of flapping wings is periodic, in most cases the movement is too fast for the system to detect its periodicity. As a result, motion presence is much more relevant to such fast movement in the retrieval process than the other properties.

The forth test is a flowing river sequence. Our system relies on image-level dynamics to retrieve similar videos of such a no-where static scene.

The last test is a video of a moving car exhibiting a rigid motion. Our system is able to separate rigid motion from non-rigid motion, but has no discrimination among rigid objects. The system retrieves vehicle sequences because they are the only rigid objects in our database.

4. CONCLUSIONS

In this paper, we propose to use higher-level motion representations to retrieve dynamic objects in videos. The motion content in terms of changes in the image context of the circumscribing box of dynamic objects is considered. Although a few assumptions have been made to implement the system, the results suggest that higher-level motion representations certainly help to retrieve a wide range of similar motion patterns. Other video contents such as color and shape are not considered in the paper, though using them will surely improve overall performance. For videos of multiple moving objects, the relationships between the corresponding circumscribing boxes can be further explored so that a query with higher-level concept such as "object A chasing object B" can be answered. However, it is out of the scope of this paper. Our main future research direction is to include other motion representations in order to cover more real-world motion patterns.

5. REFERENCES

[1] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *IEEE Workshop on Applications of Computer Vision*, 1996.

Figure 2: In each row, the leftmost image is an example image frame of the query sequence of a dynamic object. The corresponding four most similar sequences are shown in the right.

[2] M. Brand. Subspace mappings for image sequences. *Statistical Methods in Video Processing*, 2002.

[3] R. Cutler and L. Davis. Robust periodic motion and motion symmetry detection. In *Computer Vision and Pattern Recognition*, pages II:615–622, 2000.

[4] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *Computer Vision and Pattern Recognition*, pages 928–934, 1997.

[5] C. Kambhamettu, D. B. Goldgof, D. Terzopoulos, and T. S. Huang. Nonrigid motion analysis. In *Handbook of PRIP: Computer Vision*, volume 2, pages 405–430. Academic Press, 1994.

[6] J. P. Lewis. Fast normalized cross-correlation. *Vision Interface*, pages 120–123, 1995.

[7] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Tran. Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[8] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto. Dynamic texture recognition. In *Computer Vision and Pattern Recognition*, volume 2, pages 58–63, 2001.

[9] S. Soatto, G. Doretto, and Y. Wu. Dynamic textures. In *IEEE International Conference on Computer Vision*, pages 439–446, 2001.

[10] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.