SEARCHING HUMAN BEHAVIORS USING SPATIAL-TEMPORAL WORDS

Huazhong Ning, Yuxiao Hu, Thomas S. Huang

Beckman Institute and ECE Department
University of Illinois at Urbana-Champaign
405 North Mathews Avenue, Urbana, IL 61801
{hning2, hu3, huang}@ifp.uiuc.edu

ABSTRACT

This paper proposes an approach to searching human behaviors in videos using spatial-temporal words which are learnt from unlabelled data with various human behaviors through unsupervised learning. Both the query and the searched videos are represented by codewords frequencies, which capture the intrinsic information of motion and appearance of human behaviors. This representation further enables us to make use of integral histograms to accelerate the searching procedure. The performance also benefits from our feature representation that, through a MAX-like operation, may simulate the cortical equivalent of the machine-vision "window of analysis"[1]. Examples of challenging sequences with complex behaviors, including tennis and ballet, are shown.

Index Terms— Spatial-Temporal Words, Human Behavior Searching, Video Matching

1. INTRODUCTION

Searching similar human behaviors in large video database or on internet has wide applications such as video surveillance, sports video analysis, and content-based video retrieval. An intuitive idea to solve this problem is to "correlate" a short query video against the searched video sequences; the video locations with high behavioral similarities are selected as the matched positions. However, measuring similarity of natural human behaviors in video clips has proven to be very challenging for computers. One difficulty is that the same action, performed by two different people or even by the same person but at different time, are subject to large appearance variation due to different movement, scale, clothing, *etc*. Hence the searching based on unconstrained motion estimation or optical flow is highly unreliable. Although patch-based approach can alleviate this difficulty to some extent [2, 3], the computational cost is very high due to "correlation" in the 3D((x, y, t)) space. Another challenge is that, with moving cameras, nonstationary background, and moving target, few vision algorithms could identify and localize such motions well.

A lot of previous work has been presented to address these problems. Motion and trajectories are commonly used features for recognizing human actions and exhibit discriminative capability in previous work [4, 5, 6]. But estimation of optical flow or motion may be noisy due to the fundamental hurdles of aperture and singularities problems, *etc* [3]. Laptev and Lindeberg [7]'s approach, which detects a sparse set of space-time corner points to characterize the action, can partly avoid these problems. But the performance may degrade due to occlusions and misdetections of these interests points [3]. Another attempt is to measure the "behavioral similarity" by intensity/gradient on pixel level or on space-time patch level [3, 2, 8]. This method requires no foreground/background segmentation as needed in [6] and no motion segmentation. It also tolerates appearance variance in scale, orientation, and movement to some extent. Our approach, using space-time patches as well, shares these advantages.

This paper also uses patch-based feature that is the histogram of responses of a bank of 3D Gabor filters, followed by a MAX-like operation [1]. This feature is locally invariant to a range of scales and positions. Then spatial-temporal words (i.e., "bag-of-words" model) is used to represent the query video and each sliding window in the searched video. And the human behavior similarity is naturally measured by the discrepancy of codewords frequencies. "Bag-of-words" model was initially used in the text retrieval community for analyzing documents [9] and then it achieved significant success in object and natural scene categorization [10, 11]. Here the codewords dictionary is obtained by unsupervised learning from a dataset with various human behaviors.

The contributions of this work are as follows: 1) Spatial-temporal words are proposed for video representation, which not only captures the intrinsic information of motion and appearance of human behaviors, but also speeds up the scanning through integral histograms. 2) The proposed patch-based feature is locally invariant to a range of scales and positions while maintains selectivity to some extent.

2. FEATURE REPRESENTATION

Our patch-based feature is inspired by the standard model (HMAX) of object recognition in primate cortex proposed by

Riesenhuber and Poggio [1]. This feature is a histogram obtained by the following steps. Firstly the original video is convolved with a bank of 3D Gabor filters. Then we pool over limited ranges in Gabor responses through a MAX-like operation. Each Gabor orientation after pooling forms a bin of the feature histogram. Riesenhuber and Poggio [1] argued that the MAX-like operation may represent the cortical equivalent of the machine-vision "window of analysis" through which to scan and select input data. They also claimed that it is a key mechanism for object recognition in the cortex. The feature extraction is detailed as follows.

Firstly, the video frames are down-sampled to 320×240 or smaller (while maintaining the aspect ratio) to save computations. The down-sampled video is convolved with a bank of 3D Gabor filters. The Gabor filter is composed of two main components, the sinusoidal carrier and the Gaussian envelope. It exhibits many common properties, such as spatial localization, orientation selectivity and spatial frequency characterization, to mammalian cortical cells. After a minor modification of the general N-Dimension Gabor filter, we have:

$$\begin{aligned} G(x,y,t) = &\exp\left(-\left(\frac{X^2}{2\sigma_x^2}+\frac{Y^2}{2\sigma_y^2}+\frac{T^2}{2\sigma_t^2}\right)\right) \\ &\times \cos\left(\frac{2\pi}{\lambda_x}X\right)\cos\left(\frac{2\pi}{\lambda_x}Y\right) \end{aligned} \tag{1}$$

where

$$\begin{aligned} \begin{pmatrix} X \\ Y \\ T \end{pmatrix} = &\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} \\ &\times \begin{pmatrix} \cos(\omega) & 0 & \sin(\omega) \\ 0 & 1 & 0 \\ -\sin(\omega) & 0 & \cos(\omega) \end{pmatrix} \begin{pmatrix} x \\ y \\ t \end{pmatrix}. \end{aligned} \tag{2}$$

Here θ and ω are used to selectively rotate the filter at particular orientations in 3D space. Both θ and ω take discrete values $[-\frac{\pi}{4}, 0, \frac{\pi}{4}]$, so there are 9 orientations in total. Other filter parameters (the filter size, the effective width σ and the wavelength λ) are determined by considering the profiles of V1 parafoveal simple cells [12], the setup for 2D static images in [13], and the computational cost. We choose 4 scales that is formed into 2 bands, so there are 4×9 filters.

Then we pool over limited ranges in Gabor responses through a MAX-like operation. This operation takes max over grids with size $8 \times 8 \times 4$ and step size $4 \times 4 \times 2$ in each scale and then takes max over two scales of each band. After MAX-like pooling, there are 2 bands and 9 orientations per location and the video size is down sampled by $4 \times 4 \times 2$, and the feature may tolerate large variances of scales and positions.

After MAX-like pooling, the query video and each sliding window are further divided into $4 \times 4 \times 3$ patches with step $1 \times 1 \times 1$. For each patch, the responses with the same orientations are summed up so that they form two 9-bin histograms. This histogram pair is the final patch-based feature.

3. SEARCHING BY SPATIAL-TEMPORAL WORDS

With the extracted features, the spatial-temporal words (codewords dictionary) are learnt from a collection of unlabelled videos by unsupervised learning. The query video and each scanning window are represented by the frequency of spatial-temporal words. Then the query video is "correlated" with the searched video at all scanning positions. The integrate histograms are used to speed up the correlation. Figure 1 shows the flowchart of our approach.

3.1. Learning Spatial-Temporal Words

The spatial-temporal words are leant from a collection of unlabelled videos. Firstly, the features (9-bin histograms) are extracted from all patches in the original training videos according to Section 2. These features are used to train a GMM (Gaussian Mixture Model),

$$G(x) = \sum_{i=1}^{N} \alpha_i p_i(x) \tag{3}$$

where α_i's are weights and p_i's are Gaussian components. It is natural to use $\{\alpha_1 p_1, \alpha_2 p_2, \cdots, \alpha_N p_N\}$ to represent the spatial-temporal words dictionary. Intuitively, each word $\alpha_i p_i$ represents a dominant orientation response of 3D Gabor filters that is weighted by its prior probability α_i in the video collection. Note that two GMMs are trained corresponding to two bands.

The selection of the number of Gaussian components N is a trade-off between bias and variance. This paper uses Bayesian Information Criterion (BIC) [14] to select N, where the fitting of GMM model is carried out by maximization of a log-likelihood,

$$BIC = -2 \cdot loglik + (\log M) \cdot d \tag{4}$$

where $loglik$ is the log-likelihood given the samples, M is the number of samples and d is the number of parameters. In experiments we calculate BIC with N varying from 9 to 50 and $N \approx 20$ gives the maximum BIC.

3.2. Searching

Searching is carried out by correlating the query video against the searched video at all sliding windows. We extract features from each patch in the sliding window (or query video) following Section 2. Then each feature is fed into the GMM model that outputs the spatial temporal words $\{\alpha_1 p_1, \alpha_2 p_2, \cdots, \alpha_N p_N\}$. All such outputs in the sliding window (or query video) are added up together, which gives the frequencies of the codewords. These frequencies are normalized to one so that the similarity between the query video and the sliding

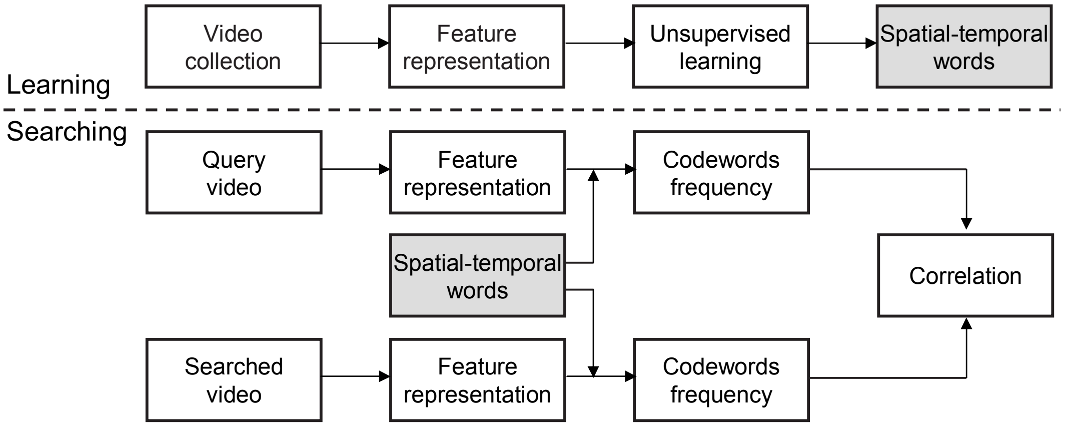

Fig. 1. **Framework.** The flowchart of our approach.

window is naturally measured by KL divergence [15]:

$$KL(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx \quad (5)$$

$$= \int f \log f dx - \int f \log g dx \quad (6)$$

where f and g are normalized frequencies of query video and sliding window respectively. To make the measure symmetric, we take the symmetric KL divergence: $d(f, g) = KL(f||g) + KL(g||f)$. The final similarity is the negative sum of KL divergences of two bands. We select the locations where the similarities are greater than a predefined threshold, but the duplicate candidates are consolidated using the neighborhood suppression algorithm from [16].

Correlation at all positions is usually very time-consuming. However, computation of the codewords for each patch can be carried out beforehand. The computational cost can be further reduced by representing the codewords frequencies in the form of integral histograms [17] so that a codeword frequency of each sliding window can be obtained by only **7** "add/subtract" operations.

4. EXPERIMENTAL RESULTS

Our feature representation and the searching using spatial-temporal words has wide applications ranging from sports video analysis, surveillance, to internet video searching. We search a short query video, which represents the human behavior of interest, in longer videos and return the occurrence of the similar behaviors. The method requires no background / foreground segmentation and tolerates a range of scales, positions and motion variations. Two experimental results are shown below: one is searching tennis strokes and the other is searching ballet turns.

Figure 2 shows the results of searching strokes in tennis videos. The short query video is a tennis stroke of 31 frames of 104×124 pixels. Figure 2 (a) shows a few frame samples. The query video is searched in a longer video playing tennis (800 frames of 228×146 pixels). We build integral histograms of codewords frequencies for the searched video and total codewords frequencies for the query video. Then the similarity between the query video and all sliding windows are computed. Figure 2 (b) shows some searched tennis strokes marked by rectangles. Figure 2 (c) draw the similarity surfaces corresponding to frames in (b), where the *yellow* indicates high similarity and *blue* the low similarity. In correlation, the query clip and the sliding window are aligned at the top-left corner instead of the center, so the peaks of the similarity surfaces do not exactly align with the people in Figure 2 and 3.

Figure 3 shows the results of searching turn actions in a ballet footage downloaded from the web ("Birmingham Royal Ballet"). It contains 400 frames of 192×144 pixels. The query video is a single turn of 20 frames with resolution 96×122. Some sample frames are shown in Figure 3 (a). Figure 3 (b) and (c) show some searched ballet turns and the corresponding similarity surfaces, respectively. Column 3 and 8 are marked as occurrences while the corresponding similarity surfaces have no salient peaks. This is because they are not the first frames of the candidate video segments and the peaks may quickly collapse after the first frame. This example is very challenging because 1) Both the query and searched videos contain fast moving parts; 2) the female dancer in the searched video wears skirt, which is different to the query video; 3) the variation in scale relative to the template is large, while our method detects most of the turns of two dancers. Shechtman and Irani [3] have tested their method on this video using the same query video. Careful comparison shows that both approaches achieve similar performance.

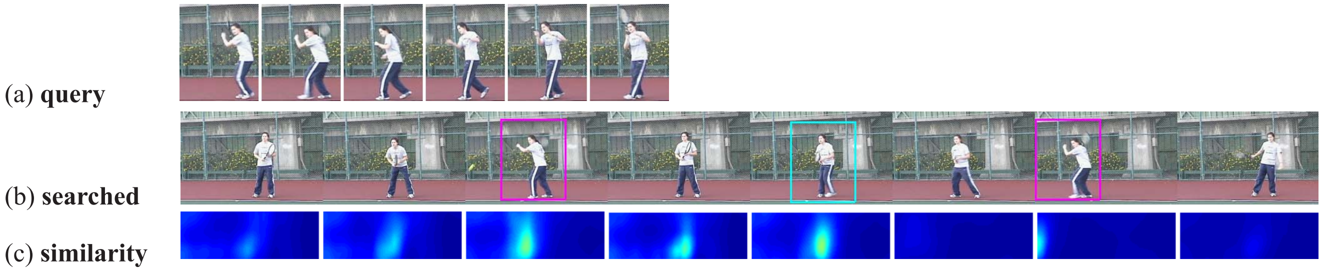

Fig. 2. **Tennis.** (a) Query video of a stoke. (b) shows some searched results and searched strokes are marked by rectangles. (c) are the similarity surfaces corresponding to the frames in (b).

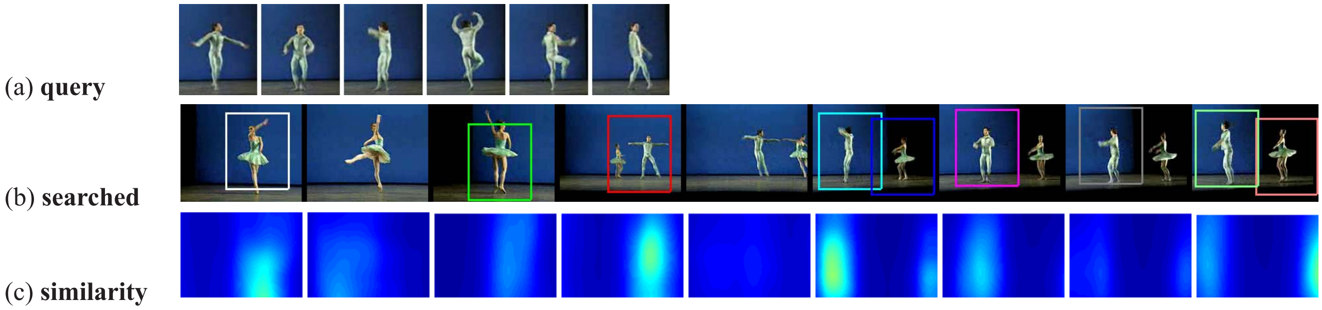

Fig. 3. **Ballet.** (a) Query video of a single turn. s(b) shows some searched results and searched strokes are marked by rectangles. (c) are the similarity surfaces corresponding to the frames in (b).

5. CONCLUSIONS

This paper presents an approach to searching human behaviors in videos using spatial-temporal words ("bag-of-words"). The contributions of this paper are twofold. One is that it learnt spatial-temporal words to represent query and searched videos that not only capture motion and appearance information but also speeds up the scanning through integral histograms. The other is that the patch-based feature is locally invariant to a range of scale and position variations while maintaining selectivity to some extent. The experiments demonstrate the effectiveness of our approach.

6. REFERENCES

[1] Maximilian Riesenhuber and Tomaso Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, 1999.

[2] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *CVPR*, 2005.

[3] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *CVPR*, 2005.

[4] M. J. Black, "Explaining optical flow events with parameterized spatio-temporal models," in *CVPR*, 1999.

[5] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik, "Recognizing action at a distance," in *ICCV*, 2003.

[6] Aaron F. Bobick and James W. Davis, "The recognition of human movement using temporal templates," *PAMI*, vol. 23, no. 3, pp. 257–267, 2001.

[7] Ivan Laptev and Tony Lindeberg, "Space-time interest points," *ICCV*, 2003.

[8] Lihi Zelnik-Manor and Michal Irani, "Event-based analysis of video.," in *CVPR*, 2001, pp. 123–130.

[9] Thomas Hofmann, "Probabilistic latent semantic indexing.," pp. 50–57, 1999.

[10] L. Fei-Fei and P. Perona, "A bayesian heirarcical model for learning natural scene categories," *Proc. CVPR*, 2005.

[11] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," *ICCV*, 2005.

[12] D. Hubel and T Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *J. Neurophysiol*, vol. 28, pp. 229–289, 1965.

[13] Thomas Serre, Lior Wolf, and Tomaso Poggio, "Object recognition with features inspired by visual cortex," in *CVPR*, 2005.

[14] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer, August 2001.

[15] S. Kullback, *Information Theory and Statistics*, Dover Publications, New York, 1968.

[16] Shivani Agarwal and Aatif Awan, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, 2004.

[17] Fatih Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *CVPR*, 2005.