A Unified Model for Activity Recognition from Video Sequences

Esther Resendiz and Narendra Ahuja
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{*eresendi, ahuja*}@*vision.ai.uiuc.edu*

Abstract

We propose an activity recognition algorithm that utilizes a unified spatial-frequency model of motion to recognize large-scale differences in action using global statistics, and subsequently distinguishes between motions with similar global statistics by spatially localizing the moving objects. We model the Fourier transforms of translating rigid objects in a video, since the Fourier domain inherently groups regions of the video with similar motion in high energy concentrations within its domain to make global motion detectable. Frequency-domain statistics can be used to isolate the frames that both adhere to our model and contain similar global motion, thus we can separate activities into broader classes based on their global motion. A least-squares solution is then solved to isolate the spatially discriminative object configurations that produce similar global motion statistics. This model provides a unified framework to form concise globally-optimal spatial and motion descriptors necessary for discriminating activities. Experimental results are demonstrated on a human activity dataset.

1. Introduction

Activity recognition is vital to many applications including surveillance and video indexing. We propose a video model that uses a parameterized approach where one assumes that a video contains moving objects, and then attempts to extract both the motion and appearance of these objects. This is useful to a variety of applications where the domain knowledge is limited and one wishes to create a concise set of intuitive features that describe how motions vary and how their spatial configurations vary. We create a frequency-domain model that allows us to discover global motion differences between extremely different activity classes (which we refer to as meta-classes), and to localize areas where discrimi-

native spatial configurations occur and solve for these local features. Human activity meta-classes could be: humans walking versus humans staying in place while performing some action. In surveillance, global pre-processing methods that isolate video frames containing locally interesting activity are useful.

The main contributions of this paper are: (**1**) a unified spatial-frequency domain model for analyzing moving objects in a video. (**2**) Using the same model, we demonstrate global discrimination of meta-classes and spatial isolation of regions that produce similar global motion but have differing local properties. Finally, (**3**) we create a generative description for activities where spatial regions which discriminate motion classes are isolated. This paper presents background, the model, and experiments to verify this approach.

2. Background

Activity recognition is a well-studied problem. Most work either derives global features from spatio-temporal gradients [8], or analyzes spatio-temporal cubic interest points detected using space-time gradients [6, 5]. Additionally, domain-specific algorithms have been developed that incorporate prior knowledge of the activity being performed [4]. The work of [8] provides global recognition of motion at a lower recognition rate than the state of the art using statistical operations of histograms, but are non-parametric and amenable to a variety of motion scenarios. Works such as [5] and [6] use spatio-temporal cubes and attempt to find cubes which are representative of a particular activity.

Although all methods achieve high levels of success on activity recognition databases, none provide a congruent method that, using one modeling framework, incorporates both global motion classification and spatial localization of moving components. Also, no other method can examine a video in its entirety and determine which frames are more immune to noise. Our work has similar advantages to many of the previous

methods. We have a global scope of video motion so our features and templates are chosen with global knowledge, since energy is concentrated in the Fourier domain along the trajectory of dominant moving objects.

3. Frequency-Domain Signal Extraction

Our model for moving objects in video uses the frequency domain, since it is widely known to have a global scope. Recent work has shown that spatial localization can be computed if motion adheres to this model [2]. We obtain the spatial Fourier transform of each video frame at time t, $I_t(\bar{\omega})$, and model it as

$$I_t(\bar{\omega}) = \sum_{l=1}^{L} S_l(\bar{\omega}) e^{j\bar{\omega}t\bar{v}_l(t)} + V_{noise}(\bar{\omega}) - V_{back}(\bar{\omega})$$

(1)

where there are L objects, each with spatial Fourier transform $S_l(\bar{\omega})$, and each with velocity $\bar{v}_l(t)$ (with respect to its position at $t = 0$). $V_{back}(\bar{\omega})$ is occluded background and $V_{noise}(\bar{\omega})$ is noise. Each frame is $N \times N$ pixels. This was described in [2]. Equation 1 models rigid objects, since $S_l(\bar{\omega})$ is time invariant.

Based on a technique called mu-propagation [3], setting $\bar{\omega} = (\mu_1, \mu_2)$, a signal $z_{\mu_1,\mu_2}(t)$ is introduced such that $z_{\mu_1,\mu_2}(t) = I_t(\mu_1, \mu_2)$, so that

$$\begin{aligned} z_{\mu_1,\mu_2}(t) &= \sum_{l=1}^{L} S_l(\mu_1, \mu_2) e^{j(\mu_1 t v_{l,x}(t) + \mu_2 t v_{l,y}(t))} \\ &+ V_{noise}(\mu_1, \mu_2) - V_{back}(\mu_1, \mu_2) \end{aligned}$$

(2)

where the x and y velocity of each object l is represented as $v_{l,x}(t)$ and $v_{l,y}(t)$ respectively. Since μ_1 and μ_2 can also be expressed as $\mu_1 = \frac{2\pi m_1}{N}$ and $\mu_2 = \frac{2\pi m_2}{N}$, where (m_1, m_2) refers to a spatial frequency bin if $I_t(\bar{\omega})$ is implemented discretely.

We implement an M-length discrete short-time Fourier transform (STFT) $Z_{\mu_1,\mu_2}(p,t)$ for the signal $z_{\mu_1,\mu_2}(t)$, and now examine each STFT frequency bin $-\frac{M}{2} \leq p \leq \frac{M}{2}$ at time t. The frequency bin p is produced by the demodulator $\omega_p = \frac{2\pi p}{M}$, such that

$$|Z_{\mu_1,\mu_2}(p,t)| = |\sum_{h=0}^{H-1} z_{\mu_1,\mu_2}(t+h) w(h) e^{-j\omega_p h}| \quad (3)$$

where $w(h)$ is a windowing function and H is the length of that window [7]. The signal $Z_{\mu_1,\mu_2}(p,t)$ is the demodulation of $z_{\mu_1,\mu_2}(t)$ by the frequency ω_p. Therefore, if $\bar{v}_l(t) = \bar{V}_l$ (constant-valued), then due to object l, $|Z_{\bar{\mu}}(\omega_p, t)| \propto |S_l(\bar{\mu})|$ in the frequency bin p that

matches the modulating velocity $\bar{V}_l$. This occurs when $\bar{\mu}\bar{V}_l = \omega_p$, which will create a peak value at frequency bin $p = \bar{m}\frac{M}{N}\bar{V}_l$. Thus, if an object l_0 travels with $\vec{V}_l$ during the time $t+1$ to $t+h$, then Equation 3 becomes

$$\begin{aligned} |Z_{\bar{\mu}}(p,t)| &= |\sum_{h=0}^{H-1} S_{l_0}(\bar{\mu}) w(h) \\ &+ \sum_{h=0}^{H-1} \left(\sum_{l=1}^{L} S_l(\bar{\mu}) \prod_{k=t+1}^{t+h} e^{j(\vec{\mu}\vec{v}_l(k))} \right) w(h) e^{-j\omega_p h}| \\ &\approx |S_{l_0}(\bar{\mu}) \sum_{h=0}^{H-1} w(h)| \end{aligned}$$

(4)

where the second term in Equation 4 is negligible if there are no other objects with velocity $\vec{V}_l$. It becomes evident that in real world applications, noise is introduced from other objects with velocities close to $\vec{V}_l$, from a time-variant $S_{l_0}(\bar{\mu})$, and from the noise terms in Equation 1. Figure 1 shows the STFT for the spatial bin $(m_1, m_2) = (0, 10)$ over time for the 'galloping' sequence of the Weizmann database [1], resized with $N = 100$. Note that the periodicity in the motion, due to the vertical oscillation of the person while galloping.
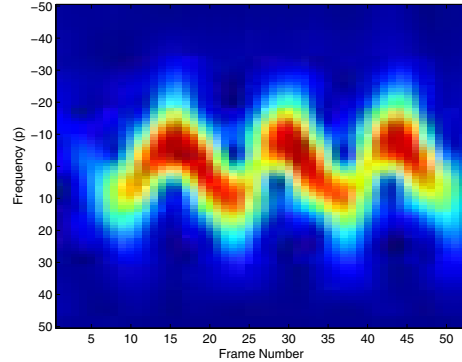

Figure 1. STFT of $z_{(0,0.628)}(t)$.

To use the STFT to both perform classification and identify consecutive frames that adhere to our model from Equation 1, we use statistics that concisely summarize the shape and the constancy of the STFT.

Spectral Centroid: The spectral centroid is a well-known function that measures the center of mass of the frequency bins as a point in time.

$$C(t) = \sum_{p=1}^{M} \frac{p|Z_{\bar{\mu}}(p,t)|}{\sum_{p=1}^{M} |Z_{\bar{\mu}}(p,t)|}$$

(5)

Entropy: We compute the entropy across the frequency bins at each time t and create the signal

$$H(t) = -\sum_{p=1}^{M} \frac{|Z_{\bar{\mu}}(p,t)|}{\sum_{p=1}^{M}|Z_{\bar{\mu}}(p,t)|} log_2\left(\frac{|Z_{\bar{\mu}}(p,t)|}{\sum_{p=1}^{M}|Z_{\bar{\mu}}(p,t)|}\right) \tag{6}$$

We then detect consecutive frames with $\frac{d}{dt}H(t) = 0$. A rigid object undergoing translational motion results in a constant entropy as an object with constant energy travels on a trajectory, as shown for the majority of frames in Figure 1. Alternately, one can look for $Z_{\bar{\mu}}(p,t)$ to be independent of time to detect consecutive frames with a constant velocity. With entropy, finding frames with $|\frac{d}{dt}H(t)| \gg 0$ is indicative of an event beyond the scope of our model (e.g. sudden appearance or disappearance of object) because a large change in entropy is indicative of a discontinuity in the phase modulation from Equation 1.

3.1 Meta-classes

The mean value as well as the amplitude of the signals $C(t)$ and $H(t)$ are used as features to differentiate global motions. One can also use periodicity, as described in [3]. We combine classes into meta-classes so that we achieve minimal error in a linear support vector machine (SVM) which uses these statistical features. This can be replaced with a domain-specific scheme.

4. Spatial-Domain Template Extraction

Once we isolate frames that adhere to our model and determine meta-class membership, the spatial domain regions are solved. Each object's displacement $\vec{b}_l$ can be determined using only the Fourier transform of the initial video frame $I(\bar{\omega})$ and the Fourier transform of a subsequent frame $I'(\bar{\omega})$ [2]

$$\begin{aligned} \frac{I'(\bar{\omega})}{I(\bar{\omega})} &= \frac{\sum_{l=1}^{L} S_l(\bar{\omega})e^{-j2\pi\bar{\omega}^T\vec{b}_l}}{\sum_{l=1}^{L} S_l(\bar{\omega})} \\ &= \sum_{l=1}^{L}\left(\frac{S_l(\bar{\omega})}{\sum_{l=1}^{L}S_l(\bar{\omega})}\right)e^{-j2\pi\bar{\omega}^T\vec{b}_l} \end{aligned} \tag{7}$$

The values $\vec{b}_l$ are then determined by peak detection after an inverse Fourier transform. The frequency domain segmentation for each object, $S_l(\bar{\omega})$, is obtained using a least-squares (LS) formulation. We construct:

$$Z = \begin{pmatrix} 1 & 1 & & \dots & 1 \\ 1 & e^{-j\omega^T\vec{b}_1} & & \dots & e^{-j\omega^T\vec{b}_M} \\ \vdots & \ddots & & \vdots & \vdots \\ 1 & e^{-j(N-1)\omega^T\vec{b}_1} & & \dots & e^{-j(N-1)\omega^T\vec{b}_M} \end{pmatrix} \tag{8}$$

and the vector $\bar{I} = [I_1(\bar{\omega})\dots I_T(\bar{\omega})]$. We solve for the frequency-based motion segmentation represented by $\bar{S} = [S_1(\bar{\omega})\dots S_L(\bar{\omega})]$ using the LS formulation

$$\bar{S} = Z^{\dagger}\bar{I} \tag{9}$$

for every frequency $\bar{\omega}$. Tikhonov regularization is used to constrain the energy of $\bar{S}$ as shown in [2].

From the L frequency-based segmentations, we obtain the $N \times N$ spatial segmentation from the inverse Fourier transform of each $S_l(\bar{\omega})$. From each spatial solution, we determine the boundaries of our object from the areas of the image with the strongest gradient. To register the spatial solution with the original image, one should look for the strongest matching gradients between $I(\bar{\omega})$ and the LS solution.

5. Experimental Results

We demonstrate the ability of our algorithm to discriminate activity and form a generative description for activity using the Weizmann database. This database contains ten actions, each performed by nine different subjects. This database contains only one object that was necessary for activity discrimination. The signal $z_{(0,0.628)}(t)$ was created according to Equation 2, and statistics were created from it. $C(t)$ corresponds to vertical motion (since m_1 is set to DC). We found that the measures $|\text{median}(C(t))|$ and $(\max(C(t)) - \min(C(t)))$ make the "stationary motions" separable from the "moving motions" of this dataset (Figure 2). We define "moving motions" as motions where a person traverses the entire screen, while in "stationary motions" the person is not traveling.

Next we discriminate the motions within each meta-class. Figure 3 shows the analysis of two subjects performing a 'galloping' motion as they move from right to left. We locate the maximum upward motion ($\max C(t)$) as in Figures 3(a-b), and then find the spatial localization. This produced the results of Figure 3 for the two subjects. We similarly analyze the maximum downward motion ($\min C(t)$). Due to human kinematics, the poses in Figure 3(c-d) are similar in appearance. The poses obtained during training are stored as templates, and during testing the correlation is measured between the test pose and the templates. We randomly separate the dataset into six training and three
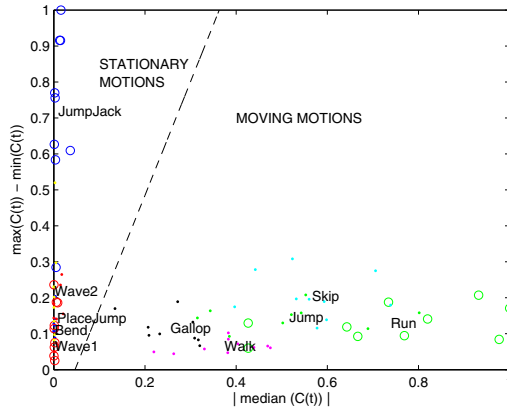
Figure 2. Meta-class classification.

testing sequences. We average the error rates of 25 experimental runs, each with a different random permutation of the dataset.
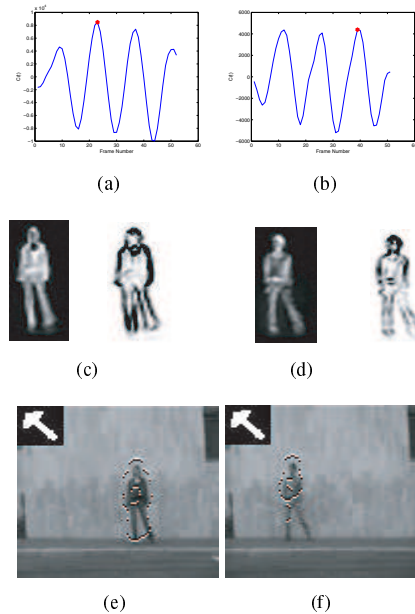

Figure 3. Galloping results for two subjects(a)-(b): $C(t)$, **(c)-(d):**$S_0(t)$, **(e)-(f): Spatial location with arrow indicating** $\vec{v}_0(t)$

This achieved an average recognition of above 80% on the database. Though this is below the state of the art in [6], the work here has several advantages. First, it forms generative model which describe *how* spatial structures differ when global motion statistics are similar. This work can be extended to more complex activities in the future. Second, the global and local information that it provides are contained within the same model so that the local information is found from globally located frames of interest. Third, it allows us to locate frames globally by summarizing all of the video's content using our statistics. Spatial templates are kept to a minimum by only being formed at both the temporally and spatially discriminative areas.

6. Conclusion

We have provided a unified model for activity recognition that utilizes the frequency domain's ability to concentrate the energy of a moving object along the velocity trajectory in the Fourier domain. This allows both a high-level categorization of motion meta-classes and a subsequent isolation of frames that discriminate the lower sub-classes. The supporting spatial regions are then identified through a least-squares solution.

For future work, we will extend our experiments to support multiple moving objects in a video. We will also develop a region-matching scheme for spatial similarity and create a similarity measure when multiple objects are present. Also, we will spatially localize the motion not just on the maximum values of $C(t)$, but on other consecutive frame where the shape of $C(t)$ is correlated across multiple videos within a class.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Action as space-time shapes. In *ICCV*, pages 1395–1402, 2005.

[2] A. Briassouli and N. Ahuja. Integrated spatial and frequency domain 2d motion segmentation and estimation. In *ICCV*, pages 244–250, 2005.

[3] A. Briassouli and N. Ahuja. Extraction and analysis of multiple periodic motions in video sequences. *IEEE Trans. PAMI*, 29(7):1244–1261, 2007.

[4] R. T. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *IEEE Intl. Conf. Auto. Face & Gesture Rec.*, pages 351–356, 2002.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Intl. Workshop on VS-PETS*, pages 65–72, 2005.

[6] J. Huang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, pages 1–8, 2007.

[7] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 2nd edition, 1999.

[8] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, pages 123–130, 2001.