DETERMINING THE SCALE OF INTEREST REGIONS IN VIDEOS

Roman Filipovych and Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory
Department of Computer Sciences
Florida Institute of Technology,
Melbourne, FL, U.S.A.

ABSTRACT

A number of action recognition methods make use of spatio-temporal features. These features often consist of local spatio-temporal descriptors centered at locations provided by an interest point detector. The extracted descriptors will then serve as input to classification algorithms. The correct scale of these descriptors is an essential parameter to be determined. Improved information quality has been achieved from recently developed entropy-based spatio-temporal feature descriptors. In this paper, we present an approach for determining scales of the sub-volumes of interest given the locations of spatio-temporal features. Our method works by measuring the average variations of local motion content calculated on subsequences of motion filter responses. We design a filter-specific data prior that allows to determine the scales of the informative neighborhoods. We demonstrate that features calculated at the scales provided by or method allow for noticeable performance improvements of action recognition algorithms.

Index Terms— Spatio-temporal Features, Action Recognition, Motion Analysis, Spatio-temporal Saliency

1. INTRODUCTION

Human action recognition has received a great amount of interest within the computer vision community in the past decade. Action recognition is a challenging problem with a number of applications including surveillance, video-retrieval, and human-computer-interaction. In this paper, we address the issue of extracting descriptive features from motion videos. Obtaining representative motion features is a crucial step in motion analysis methods. Recent action recognition methods have demonstrated the effectiveness of using local motion descriptors extracted at spatio-temporal locations across the video volume [1, 2, 3]. Once these spatio-temporal descriptors are at hand, actions can be represented as a sparse set of descriptors. This idea is largely inspired by object recognition methods [4, 5]. For example, Laptev and Lidenberg [1] extended the Harris corner detector to the spatio-temporal domain. Here, interest points are detected by analyzing spatio-temporal filter responses over increasing scales. The scale of the operator kernel determines the scale of the spatio-temporal subregion. Laptev *et al.* [2] proposes a spatio-temporal corner detector by modifying the temporal component of the scale kernel. However, the scale of the descriptor obtained from the spatio-temporal corners had a fixed uniform scale. Finally, another set of detectors are based on the adaptation of the salient region detector originally introduced by Kadir and Brady [6]. The method works by considering changes in local information content over different scales. An extension of the detector to video analysis was introduced in [3].

In this paper, we propose a method for determining the scales of detected interest regions. Rather than considering the overall change in local information content as proposed in [3], we analyze the average information change over time. In this way, we incorporate temporal variation into the scale detection process (Section 3). Additionally, we design a specific data prior that allows our method to weight higher those scales that have richer information content (Section 4). Our experiments demonstrate that our scale selection approach allows for noticeable performance improvements of action recognition algorithms (Section 5).

2. ENTROPY-BASED SALIENCY

In this section, we review the spatio-temporal salient region detector described by Oikonomopoulos *et al.* [3]. This method is a 3-D extension of the 2-D saliency detector method proposed by Kadir and Brady [6]. It begins by calculating Shannon's entropy of local image attributes inside cylindrical spatio-temporal volumes (e.g., intensity, filter response) over a range of scales:

$$\mathcal{H}_D(\mathbf{s},\mathbf{x}) = -\int_{\mathbf{q}\in D} p_D(\mathbf{q},\mathbf{s},\mathbf{x}) \log_2 p_D(\mathbf{q},\mathbf{s},\mathbf{x}) d\mathbf{q} \quad (1)$$

where $p_D(\mathbf{q},\mathbf{s},\mathbf{x})$ is the probability density function (PDF) of the signal as a function of scale $\mathbf{s}$, position $\mathbf{x}$, and descriptor $\mathbf{q}$ which takes on values from the volume D of all descriptors. The scales $\mathbf{s} = (s_1, \ldots, s_n)$ represent the size parameters of

the analyzed volumes (e.g., spatio-temporal radius cylinder's radius and length). Once the local entropy values are at hand, a set of candidate scales is selected for which the entropy $\mathcal{H}_D$ has local maxima, i.e.,

$$S = \left\{ \mathbf{s} : \frac{\partial \mathcal{H}_D(\mathbf{s}, \mathbf{x})}{\partial \mathbf{s}} = 0, \ \frac{\partial^2 \mathcal{H}_D(\mathbf{s}, \mathbf{x})}{\partial^2 \mathbf{s}} < 0 \right\} \quad (2)$$

A saliency metric $\mathcal{Y}_D$ can be obtained as a function of both scales $\mathbf{s}$ and position $\mathbf{x}$, and is defined as:

$$\mathcal{Y}_D(\mathbf{s}, \mathbf{x}) = \mathcal{H}_D(\mathbf{s}, \mathbf{x}) \mathcal{W}_D(\mathbf{s}, \mathbf{x}), \quad \forall (\mathbf{s}, \mathbf{x}) \in S \quad (3)$$

where, for candidate scales in S, the entropy values are weighted by the following interscale unpredictability measure defined via the magnitude change of the PDF as a function of scale:

$$\mathcal{W}_D(\mathbf{s}, \mathbf{x}) = \sum_i s_i \int_{\mathbf{q} \in D} \left| \frac{\partial}{\partial s_i} p_D(\mathbf{q}, \mathbf{s}, \mathbf{x}) \right| d\mathbf{q}, \quad \forall (\mathbf{s}, \mathbf{x}) \in S \quad (4)$$

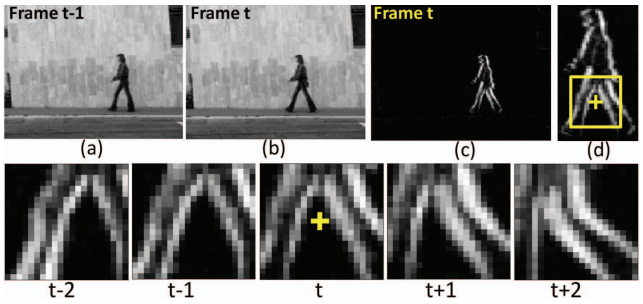

Fig. 1. Example of the motion filter and the cuboid subregion extracted around spatio-temporal point (x, y, t).

There are two main problems with the saliency measurement approach for automatic scale selection in the spatio-temporal domain. First, the saliency measurement described in [3] is based on calculating $p_D(\mathbf{q}, \mathbf{s}, \mathbf{x})$ upon all values in the spatio-temporal volume D. We propose an alternative approach that uses the average spatial entropy for saliency calculation. Secondly, all values in $p_D(\mathbf{q}, \mathbf{s}, \mathbf{x})$ are assumed to be independent and identically distributed (i.i.d., uniform prior). As the scale of the analyzing subregion increases, the zero responses (i.e., static pixels) dominate the distribution. This dominance results in a monotonically decreasing entropy, and consequently no salient region extraction can be obtained (i.e., no local maxima inside the range of the scales). Next, we propose solutions to these problems.

3. AVERAGE CHANGE IN INFORMATION CONTENT

In this section, we propose an improvement to the temporal descriptiveness of the distributions $p_D(\mathbf{q}, \mathbf{s}, \mathbf{x})$ in Equations 1 and 4. We explicitly incorporate temporal information into the entropy calculation process by representing each spatio-temporal volume by a sequence of motion filter responses $D = (D_1, \ldots, D_k)$. Motion filter responses represent instantaneous motion estimates, and can be obtained from absolute gray-level differences between two consecutive frames. Rather than considering spatio-temporal entropy, we propose to calculate the average spatial entropy on the filter response sequences as follows:

$$\widehat{\mathcal{H}}_D(\mathbf{s}, \mathbf{x}) = -\frac{1}{k} \sum_{j=1}^{k} \int_{\mathbf{q} \in D_j} p_{D_j}(\mathbf{q}, \mathbf{s}, \mathbf{x}) \log_2 p_{D_j}(\mathbf{q}, \mathbf{s}, \mathbf{x}) d\mathbf{q} \quad (5)$$

where D_j is the j-th motion filter inside spatio-temporal volume D, and k is the number of spatial subregions in D. Similarly, the average interscale saliency is defined as follows:

$$\widehat{\mathcal{W}}_D(\mathbf{s}, \mathbf{x}) = -\frac{1}{k} \sum_i s_i \sum_{j=1}^{k} \int_{\mathbf{q} \in D_j} \left| \frac{\partial}{\partial s_i} p_{D_j}(\mathbf{q}, \mathbf{s}, \mathbf{x}) \right| d\mathbf{q}, \quad \forall (\mathbf{s}, \mathbf{x}) \in S. \quad (6)$$

Figure 1 shows an example of a rectangular subregion extracted around a spatio-temporal point in a walking sequence. A sequence of 2-D subregions extracted at spatial point (x, y) in k consecutive frames constitute the video subvolume D. Next, we describe the problems arising the uniform assumption over the data, and design an appropriate data prior to address the problem.

4. FILTER-SPECIFIC DATA PRIOR

We begin by illustrating some limitations of existing spatio-temporal salient region detectors. The example consists of a hypothetical 2-D example of a motion filter response shown in Figure 2 (top-left corner). The filter response image is a vertical bar (high filter responses) surrounded by dark pixels (low filter responses). We assume that filter responses take on values from the interval $\mathbf{q} \in [0, 1]$, with $\mathbf{q} = 1$ corresponding to the maximum response, and $\mathbf{q} = 0$ when motion is absent. A location provided by an interest point detector (i.e., cross to the left of the bar) is also shown. Also, the circular region indicates the idealized scale of a local descriptor. Ideally, the saliency detector based on Equations 1 and 4 should produce a maximum at this scale (i.e., scale corresponding to the radius of the yellow circle). However, the use of a uniform prior for the calculation of $p_{D_j}(\mathbf{q}, \mathbf{s}, \mathbf{x})$ may result in a monotonically decreasing entropy, and consequently, no salient region extraction. This problem is illustrated in the first row of Figure 2. The figure shows the plots of the evolution of the entropy values, $\mathcal{H}_D$, and interscale saliency (unpredictability), $\mathcal{W}_D$, produced by Equations 1 and 4 for varying radii of the region of interest. The distribution histogram illustrates

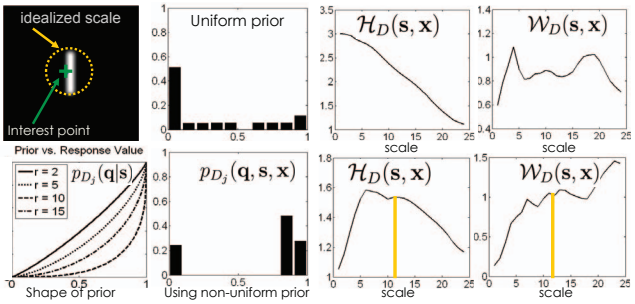

Fig. 2. Evolution of the entropy values and interscale saliency (unpredictability) using Equations 1 and 4. Left: detected interest point (green cross to the left of the bar). Dashed-line yellow circle: idealized scale.

how the zero responses (i.e., black area pixels) dominate p_{D_j} as the scale varies. Indeed, the distribution calculated using the uniform prior is highly peaked around the zero response (i.e., motionless pixels). On the other hand, the second row in Figure 2 illustrates how a non-uniform prior can produce a PDF that is able to capture richer information variation over scales. Local saliency detectors using this PDF provide better local scale estimates.

The above prior should be concentrated on high filter response values that would consequently favor the occurrence of higher filter response values over to low-response values. Our motivation is that a descriptive subregion should contain high filter response values as they indicate presence of motion in the region. We assume that locations $\mathbf{x}$ are independent of scales $\mathbf{s}$, and that both $p(\mathbf{x})$ and $p(\mathbf{s})$ are uniform. By applying Bayes' theorem, we rewrite p_{D_j} as follows:

$$p_{D_j}(\mathbf{q},\mathbf{s},\mathbf{x}) \propto p_{D_j}(\mathbf{q}|\mathbf{s},\mathbf{x}) p_{D_j}(\mathbf{q},\mathbf{s}) \tag{7}$$

Since $\mathbf{q} \in [0,1]$, we propose a prior of the form:

$$p_{D_j}(\mathbf{q},\mathbf{s}) \propto p_{D_j}(\mathbf{q}|\mathbf{s}) = \frac{1}{\mathrm{Z}}\left(1-\left(1-\mathbf{q}^{\varphi(\mathrm{V_s})}\right)^{\frac{1}{\varphi(\mathrm{V_s})}}\right) \tag{8}$$

where Z is a normalization constant, $\mathrm{V_s}$ is the volume of the subregion defined by scale $\mathbf{s}$, and $\varphi(\mathrm{V_s}) \in [1;\infty]$ is a monotonically non-decreasing real function The function $\varphi(\mathrm{V_s})$ in Equation 8 was chosen to be the following:

$$\varphi(\mathrm{V_s}) = r^{\frac{1}{2}} \tag{9}$$

where r is the spatial radius of the analyzed spatio-temporal subvolume.

5. EXPERIMENTS

Our experiments are divided into two parts. First, we provide scale selection results obtained on a set of synthetic motion filter responses. Secondly, we assess the performance of our scale selection method for action classification from videos.

We commence by demonstrating the effect of our prior distribution on automatic scale detection of simple 2-D features. Figure 3 shows a qualitative comparison between our non-uniform prior method for scale selection and the standard saliency-based method by Kadir and Brady [6]. The figure shows results on two types of simple features: a bar filter response and a corner filter response. These patterns occur frequent in filter responses obtained from human motion sequences (e.g., Figure 1). The scale obtained by standard methods using uniform prior results in non-descriptive regions. On the other hand, our non-uniform prior approach allows for better capture of relevant information.

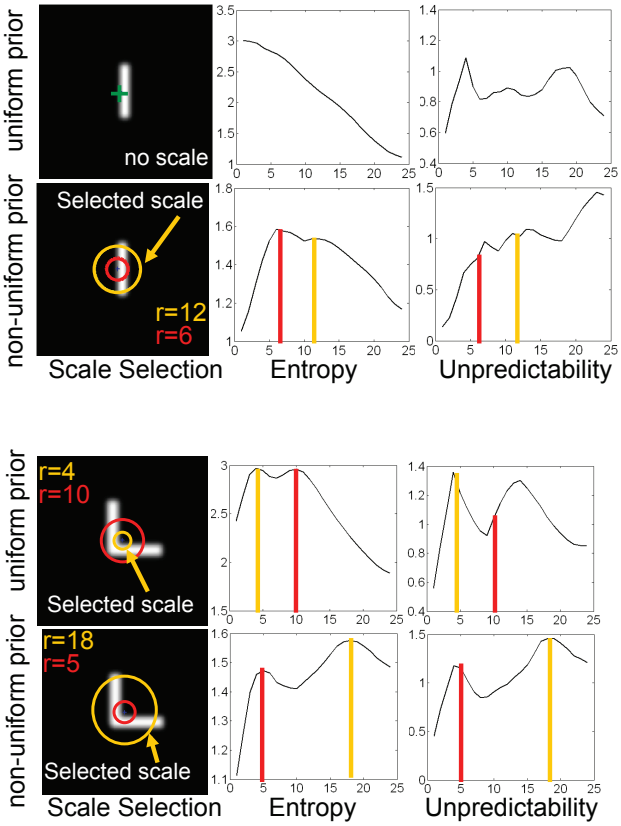

Fig. 3. Estimated subregion scales, evolution of the entropy values, and interscale saliency as produced by standard approach and using our data prior.

In the second part of our experiments, we show that existing motion classification approaches perform better when using our scale detection approach, even when detectors themselves remain intact. To accomplish this, we extract spatio-temporal subregions centered at locations provided by the detectors described in [2] and in [1]. When comparing our results with those in [2], we normalized the extracted subre-

gions to be of the same size, and extracted gradient based descriptors therein. Similarly, we used a histogram-based classification scheme, and performed evaluation on the the KTH motion dataset [7] and the Weizmann Human Action Dataset [8]. To assess our algorithm's performance using the points detected by [1], we extracted histogram of optical flow descriptors (HOF) at the detected points. We use cuboid-shaped subregions is a similar way as done in the descriptors in [1] and in [2]. We tuned the subregion scale parameters in [2] to yields the highest recognition score.

We begin by showing the effect of using the averaging in Equations 5 and 6. Figure 4 shows the relative improvements in recognition rates over [3] when using averaging and with a uniform data prior. Using this method, Equation 5 will not produce a local maxima for every interest point. Therefore, we discarded all points that did not result in a scale determination. In most cases our method improved the recognition performance. When using the detector from [2] on the Weizmann Action dataset, the scale determination using the approach from [3] marginally outperformed our method. Nevertheless, the improvements achieved using our method on interest points provided by the detector from [1] are noticable.

	KTH	Weizmann Action
Doll'ar et al.	2.5%	-1.2%

	KTH	Weizmann Action
Laptev et al.	6.2%	4.9%

Fig. 4. Relative improvement in recognition rates obtained using our equations 5 and 6 with uniform prior.

	KTH	Weizmann Action
Doll'ar et al.	4.9%	2.5%
Laptev et al.	7.4%	6.2%

(a) Relative improvement

	KTH	Weizmann Action
Doll´ar et al.	90.5%	88.1%
Laptev et al.	92.8%	91.9%

(b) % of retained points

Fig. 5. (a) Relative improvement in the recognition performance as obtained using our scale determining method on the interest locations provided by the detectors form [2] and [1], and obtained on the KTH motion dataset, and on the Weizmann Human Action Dataset. (b) Percentage of original points deemed "salient" by our method.

Next, we compared the recognition performance using features calculated using the averaging and the proposed prior. Figure 5(a) shows the relative improvements obtained without our scale determination. Figure 5(a) shows that in all cases our scale determination method allows to improve recognition rate. Finally, it is interesting to see what fraction of the original interest points was discarded as the result of our scale determination strategy. Figure 5(b) shows that usually around 90% of the points resulted in a scale determination, and were retained for classification purposes.

6. CONCLUSION

We presented an approach to accurately estimate the scale of features centered at locations provided by current spatio-temporal feature detectors. We use a saliency measure based on the average spatial entropy, and incorporates a data prior to improve local descriptor's descriptiveness. While we assumed the motion filter in the form of normalized absolute differences, extending the method to using other motion estimation techniques is straightforward. In this case, the prior should cause values corresponding to motion to have higher prior probability over scales. Future directions include to use our scale estimation method to develop our own spatio-temporal feature detector.

Acknowledgments. Research was supported by U.S. Office of Naval Research under contract: N00014-05-1-0764.

7. REFERENCES

[1] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*, Nice, France, October 2003.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, October 2005.

[3] A. Oikonomopoulos, I. Patras, and M. Pantic, "Human action recognition with spatiotemporal salient points," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 36, no. 3, pp. 710–719, 2006.

[4] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[5] Krystian Mikolajczyk and Cordelia Schmid, "Scale and affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.

[6] T. Kadir and M. Brady, "Scale saliency: a novel approach to salient feature and scale selection," in *VIE*, 2003, pp. 25–28.

[7] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: A local svm approach," in *ICPR*, 2004, vol. 3, pp. 32–36.

[8] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *ICCV*, 2005, pp. 1395–1402.