Dense saliency-based spatiotemporal feature points for action recognition

Konstantinos Rapantzikos, Yannis Avrithis, Stefanos Kollias
National Technical University of Athens
School of Electrical and Computer Engineering
{rap,iavr}@image.ntua.gr, stefanos@cs.ntua.gr

Abstract

Several spatiotemporal feature point detectors have been recently used in video analysis for action recognition. Feature points are detected using a number of measures, namely saliency, cornerness, periodicity, motion activity etc. Each of these measures is usually intensity-based and provides a different trade-off between density and informativeness. In this paper, we use saliency for feature point detection in videos and incorporate color and motion apart from intensity. Our method uses a multi-scale volumetric representation of the video and involves spatiotemporal operations at the voxel level. Saliency is computed by a global minimization process constrained by pure volumetric constraints, each of them being related to an informative visual aspect, namely spatial proximity, scale and feature similarity (intensity, color, motion). Points are selected as the extrema of the saliency response and prove to balance well between density and informativeness. We provide an intuitive view of the detected points and visual comparisons against state-of-the-art space-time detectors. Our detector outperforms them on the KTH dataset using Nearest-Neighbor classifiers and ranks among the top using different classification frameworks. Statistics and comparisons are also performed on the more difficult Hollywood Human Actions (HOHA) dataset increasing the performance compared to current published results.

1. Introduction

Given a large set of videos depicting a similar action, we often want to learn the most discriminative parts and use them to represent the action. The selection of these parts is led by small patches around *feature or interest points* based on various measures. Current methods provide sparse space-time points that are usually not enough to capture the dynamics of an action. In this work we develop a framework for volumetric saliency computation and spatiotemporal interest point detection providing a better balance between point density and discriminative power.

Although detection of spatial points has attracted the interest of many researchers, the spatiotemporal counterpart is less studied. One of the most well known space-time interest point detectors is the extension of the Harris corner detector to 3D by Laptev *et al.* in [9]. A spatio-temporal corner is defined as a region containing a spatial corner whose velocity vector is changing direction. The resulting points are sparse and roughly correspond to start and stop points of a movement when applied to action recognition. Dollár *et al.* identify the weakness of spatiotemporal corners to represent actions in certain domains (e.g. rodent behavior recognition and facial expressions) and propose a detector based on the response of Gabor filters applied both spatially and temporally. The detector produces a denser set of interest points and proves to be more representative of a wider range of actions. According to Lowe [11], sparseness is desirable to an extent, but too few features can be problematic in representing actions efficiently.

Oikonomopoulos *et al.* use a different measure and propose a spatiotemporal extension of the salient point detector of Kadir and Brady [13]. They relate the entropy of space-time regions to saliency and describe a framework to detect points of interest at their characteristic scale determined by maximizing their entropy. This detector is evaluated on a dataset of aerobic actions and promising results are reported. Wong and Cipolla in [16] report a more thorough evaluation of the latter and propose their own detector based on global information. Their detector is evaluated against the state-of-the-art in action recognition and outperforms the ones proposed by Laptev *et al.*, Dollár *et al.* and Oikonomopoulos *et al.* on standard datasets highlighting the importance of global information in space-time interest point detection. Quite recently, Willems *et al.* proposed a space-time detector based on the determinant of the 3D Hessian matrix, which is computationally efficient (use of integral videos) and is still on par with current methods.

Each of the aforementioned detectors is based on a different measure related to cornerness, entropy-based saliency, global texture or periodicity. Study of the published results confirms the trade-off, highlighted by Lowe,

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

between sparsity and discriminative power of the points. Inspired by methods related to visual attention and saliency modelling we study the incorporation of more features apart from intensity and the interaction of local and global visual information in a single measure. We derive a constrained energy formulation consisting of a data and a smoothness term, the latter containing a number of constraints. Global minimization of the energy is strongly related to figure/ground separation, since the background is continuously suppressed in each iteration.

The main contribution of our work is the computation of saliency as the solution of an energy minimization problem that involves a set of spatiotemporal constraints. Each of them is related to an informative visual aspect, namely proximity, scale and feature similarity. A secondary contribution is the incorporation of color and motion apart from intensity and global interaction in a common model. We provide an intuitive view of the resulting interest points and visual comparisons demonstrating that the detected points are dense enough to discriminate between different actions. Finally, we test the performance of the proposed model in two datasets: The KTH dataset, which is ideal for testing the repeatability and discriminative power of the points and has been often used in the literature and the HOHA dataset that is recently published and consists of a highly diverse set of movie clips. Results on the latter indicate that our detector can capture natural and real motions, not necessarily related to motion start/stop or periodicity.

2. Problem formulation

Saliency computation in video is a problem of assigning a measure of interest to each spatiotemporal visual unit. We propose a volumetric representation of the visual input where features interact to reach a saliency measure.

Figure 1 depicts a schematic diagram of the method. The input is a sequence of frames represented in our model as a volume in space-time. This volume is decomposed into a set of conspicuity features, each decomposed into multiple scales. The arrows related to the depicted pyramids correspond to voxel interactions allowed by the three constraints: (a) intra-feature (proximity), between voxels of the same feature and same scale (green arrows), (b) inter-scale (scale), between voxels of the same feature but different scale (blue arrows) and (c) inter-feature (similarity), between voxels of different features (red arrows). The stable solution of the energy minimization leads to the final saliency volume.

Let V be a volume representing a set of consequent input frames, defined on a set of points Q with $q = (x, y, t)$ being an individual space-time point. Points $q \in Q$ form a grid in the discrete Euclidean 3D space defined by their Cartesian coordinates. Under this representation point q becomes the equivalent to a voxel in this volume. Let $V(q)$ be the value

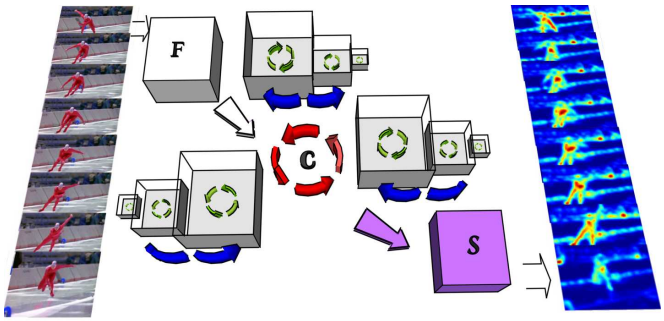

Figure 1. Schematic diagram of the proposed method (better viewed in color).

of volume V at point q .

V is decomposed into a set of conspicuity volumes C_i with $i = 1, ..., M$ corresponding to three different features, namely intensity, color and motion. Intensity and color features are based on color opponent theory and spatiotemporal orientation (motion) is computed using 3D steerable filters. Each conspicuity volume is further decomposed into multiple scales ℓ and a set $\mathbf{C} = \{C_{i,\ell}\}$ is created with $i = 1, ..., M$ and $\ell = 0, 1, ..., L$ representing a Gaussian volume pyramid.

The final saliency distribution is obtained by minimizing an energy function E composed of a data term E_d and a smoothness term E_s:

$$E(\mathbf{C}) = \lambda_d \cdot E_d(\mathbf{C}) + \lambda_s \cdot E_s(\mathbf{C}) \quad (1)$$

The data term models the interaction between the observation and the current solution, while the smoothness term is composed of the three constraints.

3. Spatiotemporal saliency and feature points

3.1. Initialization

In order to establish a common encoding and allow interaction between different features, each of the volumes participating in the energy minimization is initialized by conspicuity and not by pure feature value. Such encoding establishes a common conspicuity range among all features so that they are comparable. This means that e.g. the most conspicuous voxel in the intensity volume must have the same value as the one in the color volume.

For intensity and color, we adopt the opponent process color theory [4] that suggests the control of color perception by two opponent systems: a blue-yellow and a red-green mechanism. If volumes r, g, b are the color components of V, the intensity is obtained by

$$F_1 = \frac{1}{3} \cdot (r + g + b) \quad (2)$$

Intensity conspicuity C_1 is obtained by applying a local contrast operator to I that marks a voxel as more conspicuous when its value differs from the average value in the surrounding region:

$$C_1(q) = \left| F_1(q) - \frac{1}{|N_q|} \sum_{r \in N_q} F_1(r) \right| \quad (3)$$

where $q \in Q$ and N_q is the set of the 26-neighbors of q. The 26-neighborhood is the direct extension in 3D of the 8-neighborhood in the 2D image space.

Color conspicuity is computed as

$$C_2 = RG + BY \quad (4)$$

where $RG = |R-G|, BY = |B-Y|$ and $R = r-(g+b)/2$, $G = g-(r+b)/2$, $B = b-(r+g)/2$, $Y = (r+g)/2 - |r-g|/2 - b$.

Orientation is computed using spatiotemporal steerable filters tuned to respond to moving stimuli. The desired responses E^θ are computed by convolving the intensity volume F_1 with the second derivatives G_2 of a 3D Gaussian filter and their Hilbert transforms H_2, then taking the quadrature of the response to eliminate phase variation. More details are given in [2]. Energies are computed at orientations θ defined by the angles related to the three different spatiotemporal axis. In order to get a purer measure, the response of each filter is normalized by the sum of the consort and *orientation conspicuity* is computed by

$$C_3 = \frac{\sum_\theta E^\theta(q)}{\sum_r \sum_\theta E^\theta(r)} \quad (5)$$

All features are then decomposed into multiple scales ℓ to create the set $\mathbf{C} = \{C_{i,\ell}\}$, where $i = 1, ..., M$ and $\ell = 0, 1, ..., L$. The result is a hierarchy of video volumes that represent the input sequence in decreasing spatiotemporal scales (resolutions).

3.2. Energy formulation

The conspicuity volumes should interact in order to produce a single saliency measure for each voxel. The proposed model achieves this through a regularization framework, whereby voxels compete along a number of dimensions. Specifically, the adopted representation can be considered as a 5D one, since each voxel is connected to its 3D neighborhood at the same scale, its child/parent at a neighboring scale and to the corresponding voxel at the rest of the conspicuity volumes. Hence, we expect that voxels conspicuous enough to pop out in all dimensions will become ever salient during the minimization procedure.

The *data term*, E_d, preserves a relation between the observed and initial estimate in order to avoid excessive smoothness of the result.

$$E_d(\mathbf{C}) = \sum_i \sum_l \sum_q (C_{i,\ell}(q) - C^0_{i,\ell}(q))^2 \quad (6)$$

where $C^0_{i,\ell}(q)$ is the initial estimate, $i = 1, ..., M$, $l = 1, 2, ..., L$ and $q \in Q$. The sum limits are omitted for simplicity.

The *smoothness term*, E_s, is formulated as

$$E_s(\mathbf{C}) = E_1(\mathbf{C}) + E_2(\mathbf{C}) + E_3(\mathbf{C}) \quad (7)$$

where E_1, E_2, E_3 denote the intra-feature, inter-feature and inter-scale constraints respectively. E_1 models *intra-feature* coherency, i.e. defines the interaction among neighboring voxels of the same feature at the same scale and enhances voxels that are incoherent with their neighborhood:

$$E_1(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{|N_q|} \sum_{r \in N_q} C_{i,\ell}(r) \right)^2 \quad (8)$$

E_1 produces small spatiotemporal blobs of similar valued voxels.

E_2 models *inter-feature* coherency, i.e. it enables interaction among different features so that voxels being conspicuous across all feature volumes are grouped together and form coherent regions. It involves competition between a voxel in one feature volume and the corresponding voxels in all other feature volumes:

$$E_2(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{M-1} \sum_{j \neq i} C_{j,\ell}(q) \right)^2 \quad (9)$$

E_3 models *inter-scale* coherency among ever coarser resolutions of the input, i.e. aims to enhance voxels that are conspicuous across different pyramid scales. If a voxel retains a high value along all scales, then it should become more salient. This definition of scale saliency is in conformance with the one suggested by Kadir *et al.* [6].

$$E_3(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{L-1} \sum_{n \neq l} C_{i,n}(q) \right)^2 \quad (10)$$

3.3. Energy minimization

To minimize (1) we adopt a steepest gradient descent algorithm where the value of each feature voxel is updated along a search direction, driving the value in the direction of the estimated energy minimum

$$C^\tau_{i,\ell}(q) = C^{\tau-1}_{i,\ell}(q) + \triangle C^{\tau-1}_{i,\ell}(q) \quad (11)$$

with

$$\triangle C_{i,\ell}^{\tau-1}(q) = -\gamma \cdot \frac{\partial E(\mathbf{C}^{\tau-1}))}{\partial C_{i,\ell}^{\tau-1}(q)} + \mu \cdot \triangle C_{i,\ell}^{\tau-1}(q) \quad (12)$$

where τ is the iteration number, γ is the learning rate and μ a momentum term that controls the algorithm's stability. These two parameters are important both for stability and speed of convergence. Practically, few iterations are enough for the estimate to reach a near optimal solution. In order to keep notations simple we omit the iteration symbol τ in the following.

Equation (12) requires the computation of the energy partial derivative

$$\begin{aligned} \frac{\partial E(\mathbf{C})}{\partial C_{k,m}(s)} &= \lambda_d \cdot \frac{\partial E_d(\mathbf{C})}{\partial C_{k,m}(s)} + \lambda_d \cdot \frac{\partial E_s(\mathbf{C})}{\partial C_{k,m}(s)} \quad (13) \\ &= \lambda_d \cdot \frac{\partial E_d(\mathbf{C})}{\partial C_{k,m}(s)} + \lambda_s \cdot \sum_{c=1}^{3} \frac{\partial E_c(\mathbf{C})}{\partial C_{k,m}(s)} \end{aligned}$$

where $k = 1, ..., M$, $\ell = 1, ..., L$, $s \in Q$ and E_c with $c = 1, ..., M$ stand for the three energy constraints of the smoothness term.

In particular, the partial derivative of E_d is computed as

$$\frac{\partial E_d}{\partial C_{k,m}(s)} = 2 \cdot \sum_q \left(C_{k,m}(s) - C_{k,m}^0(s) \right) \quad (14)$$

The partial derivative of the intra-feature (proximity) constraint in (8) becomes

$$\frac{\partial E_1}{\partial C_{k,m}(s)} = 2 \cdot \left[C_{k,m}(s) - \frac{1}{|N_q|^2} \cdot \sum_{q \in N(s)} \left(2N \cdot C_{k,m}(q) - \sum_{r \in N_q} C_{i,\ell}(r) \right) \right] \quad (15)$$

The derivative of the inter-feature (similarity) constraint is computed as

$$\frac{\partial E_2}{\partial C_{k,m}(s)} = 2 \cdot \frac{M}{M-1} \cdot \left(C_{k,m}(s) - \frac{1}{M-1} \cdot \sum_{j \neq k} C_{j,m}(s) \right) \quad (16)$$

Finally, the derivative of the inter-scale (scale) constraint becomes:

$$\frac{\partial E_3}{\partial C_{k,m}(s)} = 2 \cdot \frac{L}{L-1} \cdot \left(C_{k,m}(s) - \frac{1}{L-1} \cdot \sum_{n \neq \ell} C_{k,n}(s) \right) \quad (17)$$

3.4. Interest point detection

The convergence criterion for the minimization process is defined by $max_q |\triangle C_{i,\ell}^{\tau-1}(q)| < \epsilon$, where ϵ is a small constant. The output is a set of modified conspicuity multi-scale volumes $\hat{\mathbf{C}} = \{\hat{C}_{i,\ell}\}$ and saliency is computed as the average of all volumes across features:

$$\mathbf{S} = \{S_\ell\} = \frac{1}{M} \cdot \sum_{i=1}^{M} \hat{C}_{i,\ell} \quad (18)$$

for $\ell = 1, ..., L$.

Feature points are extracted as the local maxima of the response function defined in (18). Such points are located at regions that exhibit high compactness (proximity), remain intact across scales (scale) and pop-out from their surroundings due to feature conspicuity (similarity). Hence we expect that the points will not be only located around spatio-temporal corners, but also around smoother space-time areas with distinguishing characteristics that are often important for action recognition. Visual examples and statistics in Section 4 illustrate those properties.

4. Experiments

We evaluate the proposed model by setting up experiments in the action recognition domain using two action datasets, namely the KTH dataset[1] and the Hollywood Human Actions (HOHA) one[2]. Both are public and available on-line. We provide a qualitative evaluation of the proposed detector, short descriptions of the datasets and the corresponding recognition frameworks and devote the rest of the section to quantitative analysis. For comparison purposes we use two state-of-the-art detectors, namely the periodic one proposed by Dóllar *et al.* [3] and the space-time point detector of Laptev and Lindeberg [9], which are publicly available. In the following we will denote the first one by "periodic" and the second one by "stHarris". We also compare against published state-of-the-art.

4.1. Datasets

The KTH dataset [14], one of the largest of its kind, consists of six types of human actions (handclapping, handwaving, boxing, walking, jogging and running) performed by 25 subjects in four different scenarios: outdoors (s_1), outdoors with scale variation (s_2), outdoors with different clothes (s_3) and indoors (s_4), giving rise to 2391 sequences. All sequences were recorded with a static camera at a 25fps rate and have a size of 160×120.

The HOHA dataset [10] contains video samples with human actions from 32 movies. Each sample is labeled ac-

[1] http://www.nada.kth.se/cvap/actions/

[2] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

cording to one or more of 8 action classes, namely AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp and StandUp. The training set originates from 12 movies and is manually labeled (it is the "clean" dataset in [10]). It contains 219 action samples that were manually verified to have 231 correct labels. The temporal extents of the clips were manually cropped with respect to a rough action time interval. The test set originates from 20 movies, different from the movies of the training set. It consists of 211 manually annotated action samples.

4.2. Feature point localization

In order to get an intuitive view of the proposed detector, we provide several visual examples on the KTH clips. Figure 2 shows the detected points on the 6 actions for each of the scenarios s_1 and s_2. Generally, the points are located around regions that are naturally representative of the underlying action. For example, the points at the handwaving and the boxing sequence are located on the two arms, while at the running sequence points are also detected at the legs and the torso.

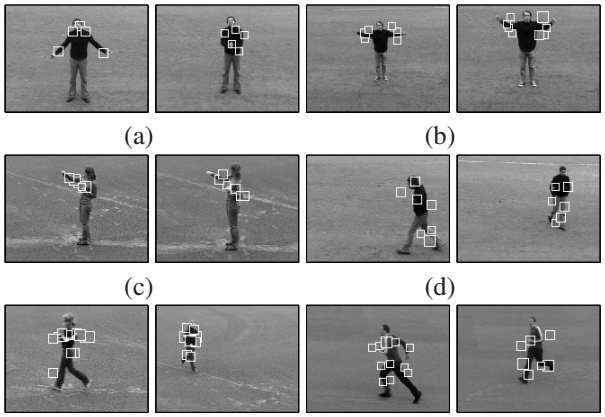

Figure 2. Results of the proposed detector on the KTH dataset. Pairs of images correspond to scenarios s_1 and s_2. (a)handwaving, (b) handclapping, (c) boxing, (d) walking, (e) jogging, (f) running

A volumetric representation of a handwaving sequence with the detected points overlaid is provided in Figure 3. Figure 3a shows 4 frames of the video. The depicted person raises his arms till the top of his head and then lowers them till the line of his waist. The result of all tested detectors are shown in Figures 3bcd. As expected, the stHarris points are located at points of velocity change, i.e. start and end points of the movement. The periodic detector focuses also at similar and neighboring regions, but is more dense. Our detector focuses both on velocity change areas and intercalary regions and therefore is able to represent smoother actions more efficiently. This property is of high importance for more complex actions, as we will see in the results

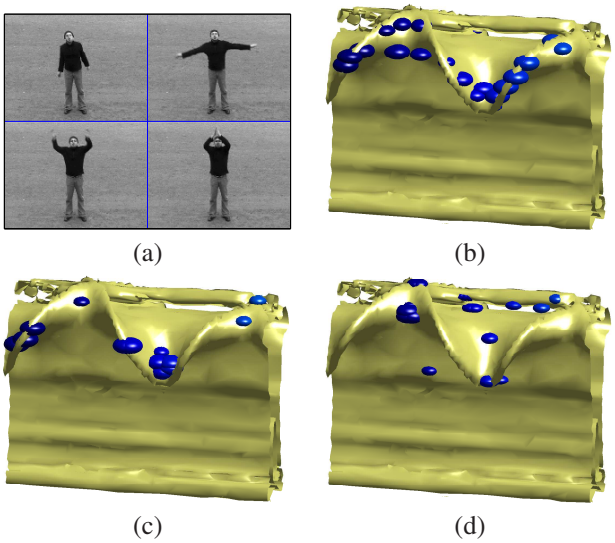

Figure 3. (a) indicative slices of a handwaving sequence and an ISO surface with the detected points overlaid for the (b) proposed, (c) periodic and (d) stHarris detectors (better viewed in color).

on the HOHA dataset. Furthermore, Figures 4, 5 depict the detected points for each of the detectors on neighboring frames of a handwaving and a walking sequence respectively. The same conclusions about the trade-off between sparsity and informativeness hold.

Method	Accuracy	Classifier
Schuldt *et al.* [14] (reported in [16])	26.95%	NNC
Schuldt *et al.* [14] (implemented by us)	50.33%	NNC
Oikonomopoulos *et al.* [13] (reported in [16])	64.79%	NNC
Wong *et al.* [16]	80.06%	NNC
Dollár *et al.* [3] (implemented by us)	79.83%	NNC
Dollár *et al.* [3]	81.20%	NNC
Ours	**88.30%**	NNC
Ke *et al.* [7]	80.90%	SVM
Schuldt *et al.* [14]	71.70%	SVM
Niebles *et al.* [12]	81.50%	pLSA
Willems *et al.* [15]	84.36%	SVM
Jiang *et al.* [5]	84.40%	LPBOOST
Laptev *et al.* [10]	91.80%	mc-SVM

Table 1. Average recognition accuracy on the KTH dataset for different classification methods. Notice that the results proposed by Laptev *et al.* in [10] are not directly comparable to ours, since the authors have used an extra optimization over a set of different descriptors

Figure 4. Cuboids detected on a handwaving sequence from KTH dataset.

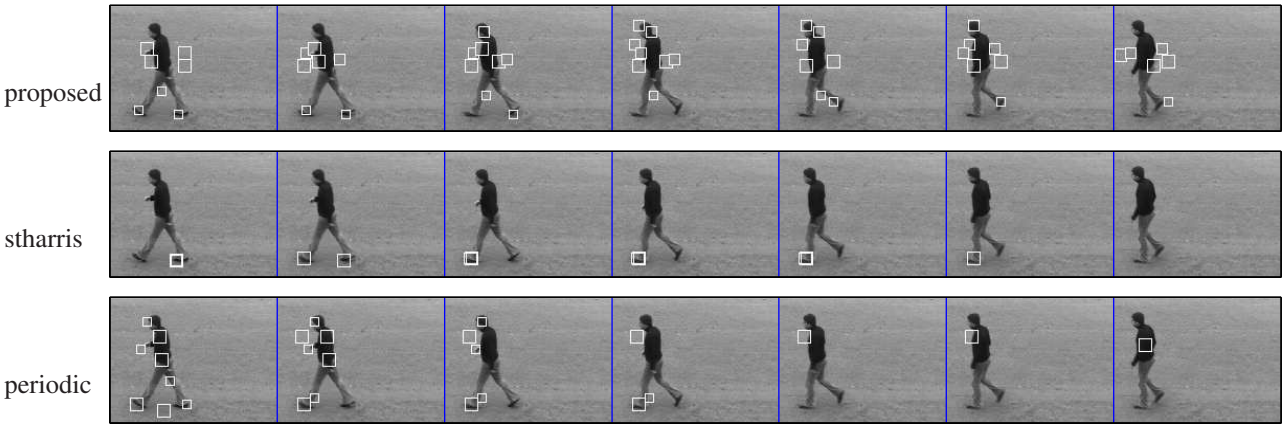

Figure 5. Cuboids detected on a walking sequence from KTH dataset.

4.3. Recognition on the KTH dataset

We use the recognition framework of Dóllar *et al.* [3] in order to provide fair comparisons and isolate the effect of the proposed detector. The same framework, which is based on bag-of-words and a Nearest-Neighbor Classifier (NNC) has been used by many researchers (see Table 1). The dataset is split into a training set and a testing set. Each training clip is represented by a set of interest points, which are detected as the local maxima of the saliency response S. A cuboid of variable size (depending on the detected scale) is defined around each point and a flattened gradient descriptor, also used by Dóllar, is extracted. We create a codebook $W = \{w_k\}, k = 1, ..., K$ of K visual words using k-means clustering on the set of descriptors from all points. The interest points of a given clip are associated to the most similar visual word and a histogram h_m of visual word occurrence is extracted for the clip.

We compute a leave-one-out estimate of the error by training on all persons minus one, testing on the remaining one and repeating the procedure for all persons. Codebooks are generated using a variable number of clusters and classification is performed using a k-NN classifier. Figures 6abc show the recognition results. Our detector performs better than the two others with an overall rate of 88.3%. It achieves rates equal to or higher than 90% for all actions except boxing. The periodic detector achieves lower rates, with the ones related to the more dynamic actions (jogging, running, walking) being higher than the rest. It seems that the inherent periodicity of these actions is well represented by the Gabor filters. The method of Laptev *et al.* ranks third with an overall rate of 50.33%.

Table 1 summarizes the results on the KTH dataset published so far. Currently, two recognition frameworks seem to be quite common, namely Nearest-Neighbor (NNC) and Support Vector Machines (SVM). Our detector achieves the best results among methods relying on NNC and is second best among all.

4.4. Recognition on the HOHA dataset

This dataset has been recently used and made public by Ivan Laptev. To our knowledge there are no published statistics except those at Laptev *et al.* [10]. Hence, we follow the same recognition framework in order to provide a reliable comparison. Nevertheless, the authors in [10] use a multi-channel recognition approach with χ^2 SVM kernel,

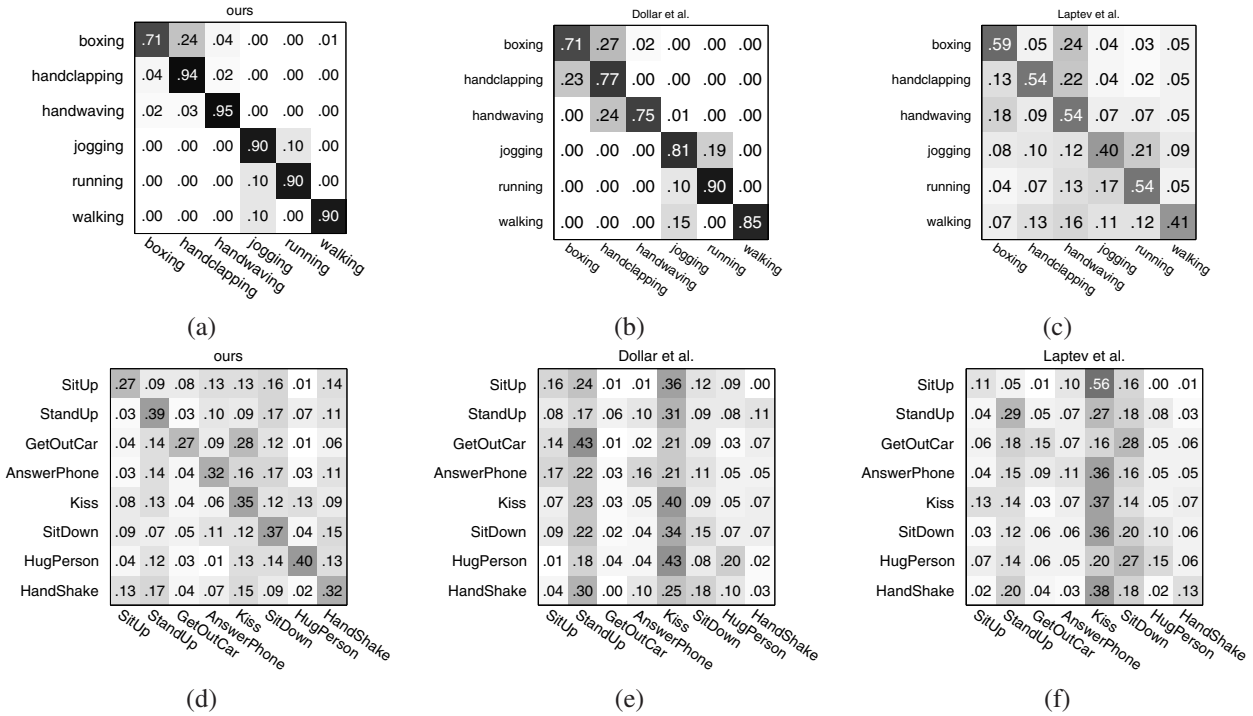

Figure 6. Confusion matrices for the KTH and HOHA datasets (row-wise).(a) our model (overall rate: 88.3%), (b) Dollár *et al.* (overall rate: 79.83%), (c) Laptev *et al.* (overall rate: 50.33%),(d) our model (500 codewords), (e) Dollár *et al.* (700 codewords) and (f) Laptev *et al.* (1300 codewords).

with each channel being the output of a trial with a different descriptor. In this work we use our own implementation with a single descriptor (flattened gradient), but we stay with the χ^2 kernel:

$$K(H_m, H_j) = \exp(-\frac{1}{A_c} D_c(H_m, H_j)) \qquad (19)$$

where $H_m = \{h_m\}$ and $H_j = \{h_j\}$ are the histograms and $D_c(H_m, H_j)$ is the χ^2 distance defined as

$$D_c(H_m, H_j) = \frac{1}{2} \sum \frac{(h_m - h_j)^2}{h_m + h_j} \qquad (20)$$

Figures 6def show the confusion matrices for the tested detectors. We experimented with different codebook sizes and report the best result for each of the detectors. As expected, the global accuracy of all detectors is low due to the diversity of the content. Nevertheless, our detector stands out and provides adequate results for all involved actions, while the performance of periodic and stHarris is not consistent across actions. Furthermore, the number of codewords needed to achieve the higher rate indicates the discriminative power of each detected feature point set. The salient points extracted by our method are dense, informative and repeatable enough so as to be represented by a shorter codebook. Figures 7 and 8 show a good and a bad detection example, respectively. Specifically, all methods successfully extract points (of different density) around the "answer-phone" action (Notice the concentration of points on the student raising his hand while holding a phone). Almost the opposite occurs in the depicted "kiss" action in Figure 8. The three girls on the right are moving quickly and most of the detected points are located on them. Since feature points are detected in a bottom-up fashion it is up to the recognition framework to select the foreground points resulting to a specific action. This is just a single bad example of a "kiss" action, while the statistics of this action are among the highest, as observed from the confusion matrices.

5. Discussion

We presented a novel spatiotemporal feature point detector, which is based on a computational model of saliency. Saliency is obtained as the solution of an energy minimization problem that is initiated by a set of volumetric feature conspicuities derived from intensity, color and motion. The energy is constrained by terms related to spatiotemporal proximity, scale and similarity and feature points are detected at the extrema of the saliency response. Background noise is automatically suppressed due to the global optimization framework and therefore the detected points are dense enough to represent well the underlying actions. We demonstrate these properties in action recognition using

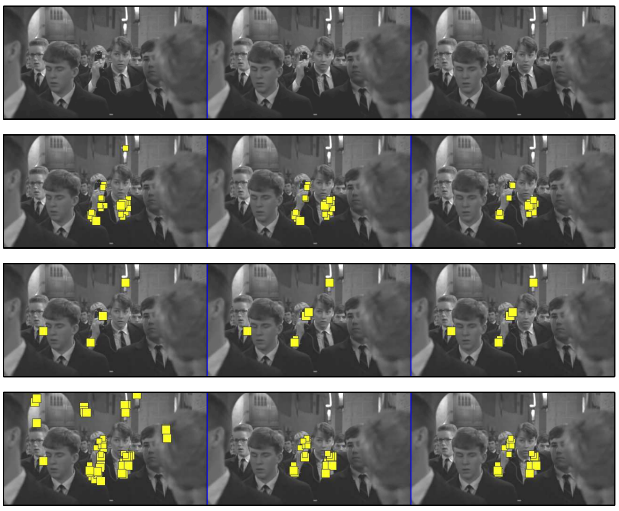

Figure 7. Detected points on an Answer-Phone action (from the clip "Dead-Poets˙1741").

Figure 8. Detected points on a Kiss action (from the clip "Butterfly˙01376").

two diverse datasets. The results reveal behavioral details of the proposed method and provide a rigorous analysis of the advantages and disadvantages of all methods involved in the comparisons. Overall, our detector performs quite well in all experiments and either outperforms the state-of-the-art techniques it is compared to or performs among the top of them depending on the adopted recognition framework. In the future, motivated by recent works, we will focus on computational efficiency issues [1] and the incorporation of advanced spatiotemporal descriptors like the ones proposed in [8].

References

[1] M. Agrawal, K. Konolige, and M. R. Blas. CenSurE: center surround extremas for realtime feature detection and matching. In *European Conference on Computer Vision (ECCV)*, volume 4, pages 102–115. Springer, 2008. 8

[2] K. G. Derpanis and J. M. Gryn. Three-dimensional nth derivative of gaussian separable steerable filters. In *IEEE International Conference on Image Processing*, 2005. 3

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72, 2005. 4, 5, 6

[4] E. Hering. *Outlines of a Theory of the Light Sense*. Harvard Univ Pr, 1964. 2

[5] H. Jiang, M. S. Drew, and Z. N. Li. Successive convex matching for action detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1646–1653, 2006. 5

[6] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001. 3

[7] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *7th Int. Workshop on Visual Surveillance*, 2007. 5

[8] A. Klaser, M. Marszalek, and C. Schmid. A Spatio-Temporal descriptor based on 3D-Gradients. pages 995–1004, 2008. 8

[9] I. Laptev. On Space-Time interest points. *International Journal of Computer Vision*, 64(2):107–123, 2005. 1, 4

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 4, 5, 6

[11] D. G. Lowe. Distinctive image features from Scale-Invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1

[12] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using Spatial-Temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008. 5

[13] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(3):710–719, 2006. 1, 5

[14] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, 2004. 4, 5

[15] G. Willems, T. Tuytelaars, and V. G. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *LECTURE NOTES IN COMPUTER SCIENCE*, pages 650–653, Marseille, France, 2008. 5

[16] S. F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007. 1, 5