# Spatio-temporal SURF for Human Action Recognition

Sameh Megrhi, Wided Souidène and Azeddine Beghdadi

L2TI, Institut Galilée, Université Paris 13, Sorbonne Paris Cité
sameh.megrhi, wided.mseddi, azeddine.beghdadi@univ-paris13.fr

**Abstract.** In this paper, we propose a new spatio-temporal descriptor called ST-SURF. The latter is based on a novel combination between the speed up robust feature and the optical flow. The Hessian detector is employed to find all interest points. To reduce the computation time, we propose a new methodology for video segmentation, in Frames Packets FPs, based on the interest points trajectory tracking. We consider only moving interest points descriptors to generate robust and powerful discriminative codebook based on K-mean clustering. We use a standard bag-of-visual-words SVM approach for action recognition. For the purpose of evaluation, the experimentations are carried out on KTH and UCF sports Datasets. It is demonstrated that the designed ST-SURF shows promising results. In fact, on KTH Dataset, the proposed method achieves an accuracy of 88.2% which is equivalent to the state-of-the-art. On the more realistic UCF sports Dataset, our method surpasses the performance of the best results of space-time descriptors/Hessian detector with 80.7%.

**Keywords:** Action recognition, SURF, optical flow, spatio-tempral feature, group of interst points, frames packets.

## 1 Introduction

In user-generated video footage, the quantity of video data containing human actions and scenes is growing exponentially (about 48 hours of video uploaded per minute on YouTubeTM) [1]. With this growth the demand for action and scene recognition or content-based video data retrieval is, certainly, colossal. Usually, considerable events are characterized by actions, for example, boxing, kissing, or some stealthy actions or behavior in a surveillance video, examples of which are shown in Fig. 1.

Recognizing human actions from videos is receiving increasing attention due to its wide range of applications such as video indexing and retrieval [2], human-computer interaction, digital entertainment, surveillance videos [3] etc. However, action recognition is usually confronted to many issues, including the necessity of handling considerable occlusions, scale changes, illumination, and the existence of background clutter, as well as viewpoint changes. In the context of action recognition in video, the representation of video objects as a bag of visual words through a histogram has become a very active research field [4]. This

Fig. 1: Samples representing three classes of human actions: boxing, kissing and running

histogram can be used in classifier framework to make the difference between object's classes. However, the main weakness of a given bag of visual words is that, not all words will be informative, accurate and objective in terms of describing actions. Consequently, the selection of the most informative words is required. The most used methods to select visual words are using Machine learning techniques Boosting [5] or adaptation process such as Multiple Instance Learning (MIL) [6] or many other State-of-the-art algorithms [7], [8]. As long as these approaches proved significant results for action recognition, they need to be adapted to be applied into the temporal domain, for action recognition or data retrieval. Recent studies in both the spatial [9] and temporal [10] domains explore the descriptive and discriminative performances of these features. In particular, spatio-temporal local features have been widely studied as image features to detect human actions, objects and events in videos. Although, video analysis with spatio-temporal features is not new, but has not been much explored yet. To extract spatio-temporal features, one of the most used methods is local cuboid. Dollar et al. [11] and Laptev et al. [12] extract Histogram of Gradient (HoG) and Histogram of Flow (HoF) from a cuboid, respectively. Although, extracting such features from a whole cuboid is not robust to noise. It is also touchy task to decide the cuboid size, and require important computational demands. In this context, we propose in this paper a novel spatio-temporal feature based on the SURF [13] local descriptor. The proposed method is based on detecting spatio-temporal interest points. We extract the descriptor by extending the original SURF descriptor to a 3D spatio-temporal Space. These descriptors are then quantized by K-means clustering and each Video clip is represented as a histogram with K bins. Support Vector Machine is then used for classification. We propose a novel codebook based on spatio-temporal descriptors called Bag of Spatio-temporal Visual Words BoSTVW. We prove experimentally that this contribution outperforms other state-of the-art approaches on the increasingly complex and popular KTH Dataset [14] and UCF sports Dataset [15].
The paper is structured as follows. Initially, an overview of recent related work is given in section 2, while section 3 explains the proposed approach. Extensive results and conclusions are presented in sections 4 and 5 respectively.

## 2 Related works

Inspired from the text retrieval community, the 'bag of words' BoW, has recently become popular for image [16] and video analysis [17]. In action recognition's

state-of-the-art, the BoW models were widely used since they have shown the effectiveness of local appearance based descriptors [18], [19]. However, in comparison with other approaches, Bag of visual word selection is still in its infancy. To extract video descriptors, many researchers have been investigating in tracking major parts of human bodies then extracting features from these regions [20]. However, they need to setup many hypothesis. These considerations and hypothesis are often demanding. So that, methods based on spatio-temporal features are promising for action recognition. Some of them are based on the extraction of low-level optical flows from cuboids [21] this method gives good results in terms of feature selection and a good classiffications accuracy [21]. But they presents limits concerning the long computational time they require. Dollar et al. detect local cuboids to apply 1-D Gabor filters in the temporal direction and 2-D Gaussian kernels in the spatial space [11], and they produce video visual words based on vector-quantizing in the same way as bag-of-visual-words for object recognition [16]. In the same direction, Laptev et al proposed STIP (Spatio-Time Interest Points) to detect cuboids [12]. This method is considered as an extension of Harris detector. Nevertheless, the limits of the aforementioned methods not only concerns the hardness of finding the best cuboid size, but also the high computational requirements. To overcome these problems, we propose to detect interest points using SURF/Hessian [13]. Then we segment the videos into Groups of interest points (GIP) and Frame Packets (FPs) to reduce the computation time. We use Sun, D at al. [22] optical flow detection methods which allows to extract spatio-temporal SURF by tracking interest points instead of cuboids.

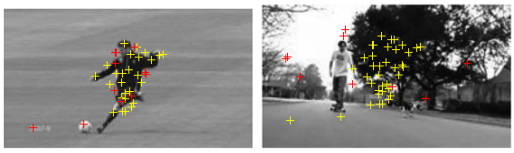

Fig. 2: Example of SURFs found using Hessian detectors on diffrent frames from UCF sports dataset.

## 3 The proposed approach for action recognition

The proposed method aims at detecting human actions, to attend this purpose, first video sequences are segmented in Frame Packets (FPs) and Group of Interest Points (GIP). Second, based on a novel combination between optical flow computed by [22] and **S**patio-**T**emporal SURF (Speeded-Up Robust Feature) [13], the interest points **ST-SURF** are localized and extracted, from all training video FPs. Then, the extracted ST-SURFs are clustered using K-means clustering algorithm. The video clips are represented as a K-bins histogram of the

quantized descriptors 'bag of spatiotemporal visual words' BoSTVW. Finally, an SVM classfier is trained using these histograms *(One vs all)*.

### 3.1 Frame Packets (FPs) and Group of Interest Points (GIP) segmentation

To be able to achieve accurate and fast computation, our algorithm does not use all the frames available in a video in order to extract its descriptors. Instead, we have created and used the concepts of Frame Packets (FPs) and Group of Interest Points (GIP). We assume that, between three successive frames (n-1, n and n+1), an interest point (from one picture to another) can have three possible states: still, moving and disappear. The first and last states are obvious because in the first case no motion is detected. In the last case the IP has disappeared and cannot be tracked any more. The second state is the one that concerns us the most, since there is a displacement and we can track the motion angle. From now on, we assume that $\alpha$ is the angle between the lines segments supporting the motion of an IP from the couple of frames (n-1, n) and (n, n+1), Fig 3. By

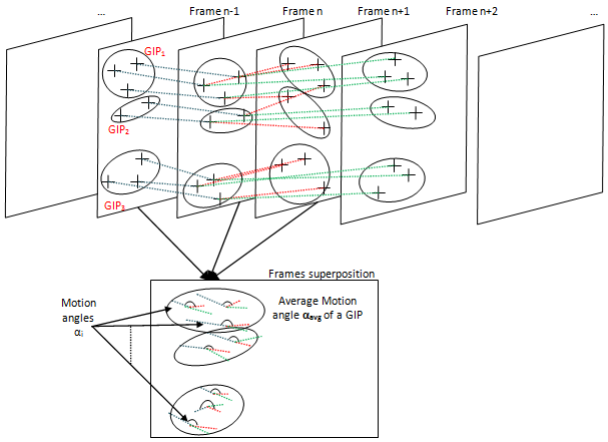


Fig. 3: IPs trajectory tracking for FPs segmentation

comparing $\alpha$ to $\alpha_{max}$ (a parameter fixed at the beginning of the processing) we are able to segment a succession of frames, that we call here Frame Packets (FPs), in which each IPs $\alpha$ is lower than $\alpha_{max}$. By calibrating this angle of tolerance, we are able to certify that, within this FP, all IPs movements are within this tolerance parameter. This means that we cannot miss any significant movement likely to influence the remaining computing. We introduce, then, the concept of Group of Interest Points (GIP) in order to be able to have more control over the size (in number of frames) of the FP. In fact, a GIP is a parameter defining the number of IP that must be grouped together. This grouping is performed over

successive IPs in a frame. By defining this number NGIP we can compute an average angle ($\alpha_{avg}$) for a certain GIP and compare it to the $\alpha_{max}$. The higher NGIP is the less the $\alpha_{avg}$ will be sensitive to motion and the more the FP will contain frames. Here are the steps of our segmentation algorithm. Let us suppose that we are beginning the computation of a new FP:

- We extract the IP of the frames one and two.
- We define the GIPs based on the NGIP parameter fixed at the beginning of the algorithm.
- We compute the line supporting the motion for each corresponding IP within these two frames.
- We apply the above three steps to the frames two and three.
- We compute the angle between each motion line and we extract the average angle for each GIP.
- We compare each average angle to the $\alpha_{max}$ (fixed at the beginning of the algorithm).
- We continue performing the above six steps over the next frames (taking, always, the first motion direction as reference to all remaining comparisons) until finding an average angle of a GIP higher than the maximum angle. In this case we can define the FP and assume, with confidence, that the first and last frames of this FP can fully describe the motion within.

### 3.2 Interest points extraction

In the following, we present the used interest point detector followed by a description of the feature we extract. For interest points detection we choose the Hessian detector [23]. It searches for image locations that exhibit strong derivatives in two orthogonal directions. It is based on the matrix of second derivatives, the so-called Hessian [23]. In 2004, Lowe [13], presented SIFT for extracting invariant features from images that can be robust against image scale and rotation. Then it was widely used in image, recognition and retrieval etc. However, extracting robust features approaches are very slow. Bay et al. speeded up robust features by using integral images for image convolutions and Fast-Hessian detector [13]. Their experiments turned out that SURF was faster and it works well. We use the extraction solution given by [13] to extract interest point feature. This choice is motivated by the robustness, the smaller size of this feature and their excellent performances attested in various datasets for action recognition [24]. The SURF feature is a 64-D vectors that describes spatial patterns around detected points. Refer to [13]for the detail.

### 3.3 Surf Tracking Into 3D Feature Space

Features' tracking is performed by estimating optical flow. To increase optical flow estimation accuracy, many researches are inspired from the Horn and Schunck (HS) Optical flow formulation [22]. In fact, they focuses on optimizing an objective function which combines the image's properties and its spatial motion

prediction. Sun end al. proposed a new algorithm to approximate an optimized computationally tractable objective function, based on the original HS formulation. They first, use median filtering to denoise the flow, Exploiting connections between median filtering and L1-based denoising. They proved that algorithms relying on a median filtering step are approximately optimizing a different objective that regularizes the flow over a large spatial neighbourhood [22]. The resulting algorithm ranks 1st in both angular and end-point errors in the Middlebury evaluation in March 2010 [22]. In our work, we considered every Frame Packet as a volume of frames in the 3D space called FP Volume ($FPV$), this cubic volume is characterized by its frames' number($FN$) , its frames' surfaces dimensions ($FS$) and its center ($FPVc$). A given interest point $IP = (x, y, t)$ is defined by its position $(x, y)$ and its frame $t$. In frame $(t + n)$, the $IP$ moves by a displacements $u$ in the $x$ direction, and $v$ in the $y$ direction. $IP$ becomes, $IP(t+n) = (x+u, y+v, t+n)$. In all our experiments, unless mentioned otherwise, we assume that due to the video segmentation into FPs, the motion vectors trajectory remain stable. For stagnant interest points $u = v = 0$. Thus, in the $FPV$, the 3D direction $(u, v, n)$ represent the direction of the $IP$ motion. The motion vector is calculated by the Sun et al. [37] optical flow approach. Our contribution consists on the use of motion orientation and position to characterize the motions, instead of using the direction vector $(u, v, n)$ generated from optical flow computation. We suppose that the motion vector in the 3D space can be defined as the intersection of two planes perpendicular respectively to the plane $(t, x)$ and the plane $(t, y)$. This parameterization is one among several possible representations of 3D lines [25]. To extract $IP$ orientation, we project its motion vectors onto the planes $(t, x)$ and $(t, y)$ of the $FPV$ to define an angle for each projection, the first angle $\alpha_x$ between optical flow and the plane $(t, x)$, the angle $\alpha_y$ between the plane $(t, y)$ and the motion vector.

$$\propto_x = 90 - \frac{180}{\Pi} arctan(u), \propto_y = 90 - \frac{180}{\Pi} arctan(v). \tag{1}$$

For each $IP$, we project its motion vector onto the planes $(t, x)$ and $(t, y)$ and obtain two lines $L_x$ and $L_y$. The orthogonal projection of $FPVc_x$ and $FPVc_y$ onto the lines $L_x$ and $L_y$ allows the computing of both distances $D_x$ and $D_y$ between the cube center and the lines supporting the motion vectors ($L_x$ and $L_y$).
For an $IP$ located at $(x, y, t)$:

$$D_x = D_{xu} - D_{tv}, D_y = D_{yv} - D_{tu} \tag{2}$$

where

$$D_{xu} = (x - x_{max}/2) cos(180/\Pi arctan(u)) \tag{3}$$

$$D_{tv} = (t - t_{max}/2) sin(180/\Pi arctan(v)) \tag{4}$$

$$D_{yv} = (y - y_{max}/2) cos(180/\Pi arctan(v)) \tag{5}$$

$$D_{tu} = (t - t_{max}/2)sin(180/\Pi arctan(u)) \quad (6)$$

where $t_{max}$, $x_{max}$ and $y_{max}$ are the dimensions of the Frame Packet volume. In the following, $D_x$ and $D_y$ describe the motion distances of a given interest point. Fig. 4, is a graphical illustration of the cube center and its projection into the planes $(t, x)$ and $(t, y)$.

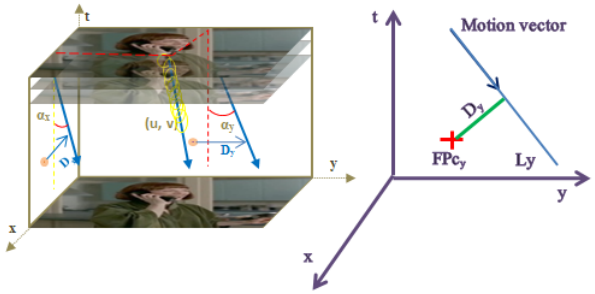


Fig. 4: The projection of a motion vector in the adjacents planes.

**ST-SURF Extraction:** This step consists in the generation of the novel ST-SURF that we designed. This descriptor is represented by spatial feature 64-D vector, and temporal 4-D feature, we concatenate both vectors into one 68-D spatiotemporal descriptor vector, and thus we extend the image SURF descriptor [13] to videos. These features will tracks the interest point through time in each FP we defined. The size of a FP depends on its average frames number. In our work we consider only moving interest points (where $\propto_x \neq 0$ and $\propto_y \neq 0$).

**ST-SURF Training Pipeline:** For the purpose of action recognition, we follow the same steps of the case of object categorization. After the extraction of ST-SURF descriptors, we define a spatio-temporal words dictionary. The basic idea is to assign a set of objects into groups so that the objects of similar type will be in one cluster, in order to construct a visual codebook, which can be used to represent an action, a scene or en object. Recently, the K-means algorithm has been widely used to construct the visual codebook because of its high performances and simplicity. A codebook is learned to quantize input features into visual spatiotemporal codewords. Fig. 5, illustrate the training steps of UFC sports dataset's videos.

**Evaluation Pipeline:** After the extraction step, the generated ST-SURFs are quantized into visual words using k-means clustering. Each video sequence can

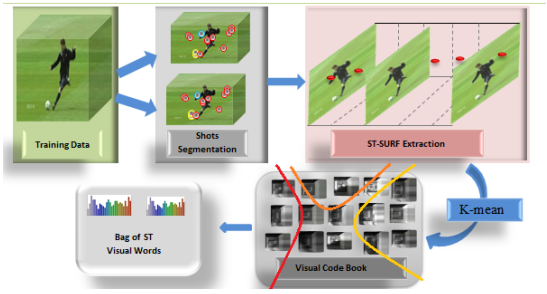


Fig. 5: Training pipeline

then be represented as the frequency histogram over the visual words. Generally, using a large-sized codebook allows to obtain high recognition accuracy, yet an oversized codebook leads to high quantization errors. The resulting histograms of visual word occurrences are used as classification inputs. We use a non linear support vector machine to classify human actions.

## 4 Experiments

In the following we describe the datasets used for the evaluation of the proposed work. We evaluate the ST-SURF in a bag-of-features based action classification task and compare our approach to the state-of-the-art employ.

### 4.1 Experimental Setups and Data

**Dataset:** The proposed framework is tested on the KTH dataset [14] and UCF sports Dataset [15]. The KTH dataset is commonly used as a public benchmark test of spatio-temporal features [24]. This dataset contains six kinds of actions such as walking, running, jogging, boxing, hand waving and hand clapping. We consider 6 action classes by 25 persons in 4 different scenarios with a total of 2391 video samples. The average length of videos in the KTH dataset is about 20 second long. The second one is the UCF sports dataset, more realistic and challenging data obtained from broadcast sport videos by Ahmed et al. [15]. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. The publicly available part of this dataset contains nine actions namely diving, golf, swinging, kicking, lifting, horseback riding, running, skating, swinging and walking. This dataset contains close to 200 video sequences at a resolution of 720x480 [15].

**Parameter Settings:** In all our experiments, we explored optimal parameter settings. We evaluate the classification rates of both KTH and UCF datasets while changing the codebook and the FPs sizes. The results shows that the empirically optimal size book is k = 4000 with $\alpha_{max} = 42\circ$ and GIP = 38. These settings gave us empirically satisfactory results.

### 4.2 Experimental results

**KTH Dataset:** From the recently reported results of the state-of-the-art, we can clearly conclude that using Hessian detector, Laptev and al. [10] obtained 88.7% using a combination of HOG (histograms of gradient orientations) and HOF (histograms of optical flow) descriptors, 88.6% using HOF, and 77.7% with HOG . We note that Kläser and al. achieved an accuracy of 84.6% using HOG3D descriptor, which is a comparable results with HOG/HOF [10]. The combination SURF/Hessian detector gives 84.6% for williems and 86% for Noguchi [24]. In the table. 1, the first row compares the **b**est **a**verage **a**ccuracy (**BAA**) for the differents detector/descriptor combinations reported by other researchers, on the KTH dataset. The **a**verage **a**ccuracy for **H**essian (**HAA**) detector/descriptor combinations on the KTH dataset are drawn in the second row.

**UCF sports Dataset:** We note that Kläser and al. achieved an accuracy of 85% using HOG3D/Gabor descriptor, Laptev and al. [10] obtain 81.6%using HOG/HOF, 82.6% using HOF, and 77.4% with HOG. The first row, in Table. 2, compares the **b**est **a**verage **a**ccuracy (**BAA**) for the differents detector/descriptor combinations reported by other researchers, on the UCF sports dataset. The **a**verage **a**ccuracy for **H**essian (**HAA**) detector/descriptor combinations on the UCF sports dataset are drawn in the second row.

Table 1: Average accuracy for various detector/descriptor combinations on the KTH dataset.

| . | HOG3D | HOG/HOF | HOG | HOF | E-SURF | t-SURF | ST-SURF |
|---|---|---|---|---|---|---|---|
| **BAA** | 90% | 91.8% | 82.3% | 92.1% | 81.4% | 86% | **88.2%** |
| **HAA** | 84.6% | 88.7% | 77.7% | 88.6 % | 81.4% | 86% | **88.2%** |

Table 2: Average accuracy for various detector/descriptor combinations on the UCF sports dataset.

| . | HOG3D | HOG/HOF | HOG | HOF | E-SURF | t-SURF | ST-SURF |
|---|---|---|---|---|---|---|---|
| **BAA** | 85% | 81.6% | 77.4% | 82.6% | 77.3% | - | **80.7%** |
| **HAA** | 78.9% | 79.3% | 66.0% | 75.3% | 77.3% | - | **80.7%** |

Tables 3-4 are the confusion matrices of the actions classification results based on two type of features. the first matrix describle the classification result for the visual SURF feature reported in [24]. The result among a single visual feature give bad classification results for all the actions. Based on Table. 4, confusion

Table 3: SURF confusion matrix action recognition on the KTH dataset.

| KTH | Boxing | clapping | Waving | Walking | Jogging | Running |
|---|---|---|---|---|---|---|
| Boxing | 0.6 | 0.01 | 0.03 | 0.13 | 0.1 | 0.1 |
| clapping | 0.06 | 0.58 | 0.25 | 0.04 | 0.02 | 0.05 |
| Waving | 0.05 | 0.08 | 0.74 | 0.03 | 0.01 | 0.09 |
| Walking | 0.01 | 0 | 0 | 0.7 | 0.16 | 0.13 |
| Jogging | 0 | 0.01 | 0 | 0.12 | 0.58 | 0.29 |
| Running | 0 | 0 | 0 | 0.1 | 0.21 | 0.59 |

Table 4: ST-SURF confusion matrix action recognition on the KTH dataset.

| KTH | Boxing | clapping | Waving | Walking | Jogging | Running |
|---|---|---|---|---|---|---|
| Boxing | 0.9 | 0.07 | 0 | 0.03 | 0 | 0 |
| clapping | 0. 07 | 0.9 | 0.03 | 0 | 0 | 0 |
| Waving | 0.01 | 0.06 | 0.93 | 0 | 0 | 0 |
| Walking | 0 | 0 | 0 | 0.91 | 0.06 | 0.3 |
| Jogging | 0 | 0 | 0 | 0.04 | 0.85 | 0.1 |
| Running | 0 | 0 | 0 | 0.06 | 0.14 | 0.8 |

matrix show that the original combination, that we proposed, of both visual and motion features (ST-SURF) boosted significantly the classification accuarcy. Regarding the average result over each of the six actions' KTH dataset, ST-SURF produced good result, however, less accuracy is observed in the jogging and running actions because thes actions are almost similar. lastly but not least, comparing with results driven by the best result of the state-of-the-art, our method achieve 88.2% better than the 86% reported by Noguchi and al. [24] using Spatio-temporal SURF. Outperforming the results of the Cuboids/HOG combinaison obtained by [12] 82.3% and the 81.4% reported by Willem and al. [26]. Based on the confusion matrix of UCF sports Dataset given in Table. 2, the ST-SURF outperform the best result driven by the state-of-the-art using Hessian detector and achieves 80.7% of accuracy. We note that ST-SURf/Hessian gave better results in realistic videos. We are still below the results driven by Laptev and al. with 85% using the HOG3D/Gabor combination, and their 91.8% reached using Harris3D/(HOG/HOF), this can be due to different codebook generation and the use of different interest points detectors. This motivates further investigations of different interest points detectors and realistic vidoe settings. Regarding all these results our method is equivalent to the state of the art, and shows significantly better performance, outperforming many results driven in the same setup.

Table 5: ST-SURF confusion matrix action recognition on the UCF sports dataset.

| UCF | Dive | Golf | Kick | Lift | Ride | Run | Skate | Swing | Walk |
|---|---|---|---|---|---|---|---|---|---|
| Dive | 0.8 | 0.17 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0 |
| Golf | 0 | 0.78 | 0.2 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| Kick | 0 | 0 | 0.9 | 0 | 0 | 0.07 | 0 | 0 | 0.03 |
| Lift | 0 | 0 | 0 | 0.92 | 0 | 0 | 0 | 0.08 | 0 |
| Ride | 0 | 0 | 0.2 | 0 | 0.62 | 0.18 | 0 | 0 | 0 |
| Run | 0 | 0 | 0.02 | 0 | 0 | 0.88 | 0 | 0 | 0 .1 |
| Skate | 0 | 0 | 0.08 | 0 | 0 | 0 | 0.6 | 0 | 0.32 |
| Swing | 0 | 0 | 0 | 0 | 0 | 0 | 0.21 | 0.79 | 0 |
| Walk | 0 | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0.96 |

## 5 Conclusion

In this paper, We have investigated a novel scheme to efficiently segment video sequences into a new concept we called Fram Packets. Then we proposed a novel spatio-temporal descriptor based spatio-temporal interest points. The designed descriptor is an extension of the SURF to the temporal domain. The proposed feature extraction consists on detecting of the Surf points and mapping them into a 3D feature space based on an original exploitation of the optical flow orientation and position. Only the moving SURF are then selected. The extracted features are embedded into a bag of visual word pipeline, to finally classify six actions from KTH Dataset and then nine actions from the UCF sport dataset. Furthermore, the proposed framework demonstrate promising recognition performance on tow standard benchmarks with the accuracy about 88.2% in KTH and 80.7% in UCF sports. We consider many perspectives for the future, the most important includes applying the same work to tackl more complex and realistic actions. We also plan to improve our ST-SURF and consider combining it with variety of low level features and use these data to video searching and retrieval. Furthermore, we look to investigate diffeterent interest points detectors and larger datasets in bag-of-visual-words based representations. Finally, the results we obtained demonstrate the viability of our approach and proove that even without refinements we already are equivalent to the state-of-the-art performances.

## References


1. Youtube: Statistiques@ONLINE (June 2009)
2. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of the 6th ACM international conference on Image and video retrieval. CIVR, New York, NY, USA, ACM (2007) 494–501
3. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on (2004) 334–352
4. Willamowski, J., Arregui, D., Csurka, G., Dance, C.R., Fan, L.: Categorizing nine visual classes using local appearance descriptors. illumination (2004) 21
5. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,CVPR., IEEE (2001) I–511
6. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems (1998) 570–576
7. Kim, T.K., Wong, S.F., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR., IEEE (2007) 1–8
8. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: IEEE 12th International Conference on Computer Vision, IEEE (2009) 444–451

9. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International journal of computer vision **65**(1-2) (2005) 43–72
10. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., et al.: Evaluation of local spatio-temporal features for action recognition. In: BMVC British Machine Vision Conference. (2009)
11. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance., IEEE (2005) 65–72
12. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: Spatial Coherence for Visual Motion Analysis. Springer (2006) 91–103
13. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer Vision–ECCV. Springer (2006) 404–417
14. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR., IEEE (2004) 32–36
15. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
16. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. (2004) 22
17. Brandão Lopes, A.P., Alves do Valle Jr, E., Marques de Almeida, J., Albuquerque de Araújo, A.: Action recognition in videos: from motion capture labs to the web. (2010)
18. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision (3) (2008) 299–318
19. Riemenschneider, H., Donoser, M., Bischof, H.: Bag of optical flow volumes for image sequence recognition. In: British Machine Vision Conf. Number 3 (2009) 4
20. Mojarrad, M., Dezfouli, M.A., Rahmani, A.M.: Feature extraction of human body composition in images by segmentation method. World Academy of Science, Engineering and Technology (2008)
21. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR., IEEE (2008) 1–8
22. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: IEEE Conference onComputer Vision and Pattern Recognition CVPR., IEEE (2010) 2432–2439
23. Beaudet, P.R.: Rotationally invariant image operators. In: Proceedings of the International Joint Conference on Pattern Recognition. (1978) 579–583
24. Noguchi, A., Yanai, K.: A surf-based spatio-temporal feature for feature-fusion-based action recognition. In: Trends and Topics in Computer Vision. Springer (2012) 153–167
25. Dementhon, D., Doermann, D.: Video retrieval of near-duplicates using $\kappa$-nearest neighbor retrieval of spatio-temporal descriptors. Multimedia Tools and Applications (3) (2006) 229–253
26. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Computer Vision–ECCV. Springer (2008) 650–663