Human Action Recognition in Videos: A comparative evaluation of the classical and velocity adaptation space-time interest points techniques

Ana Paula G S de Almeida
University of Brasilia, Brazil
anapaula.gsa@gmail.com

Bruno Luiggi M. Espinoza
University of Brasilia, Brazil
bruno@cic.unb.br

Flavio de Barros Vidal
University of Brasilia, Brazil
fbvidal@unb.br

Abstract

Human action recognition is a topic widely studied over time, using numerous techniques and methods to solve a fundamental problem in automatic video analysis. Basically, a traditional human action recognition system collects video frames of human activities, extracts the desired features of each human skeleton and classify them to distinguish human gesture. However, almost all of these approaches roll out the space-time information of the recognition process. In this paper we present a novel use of an existing state-of-the-art space-time technique, the Space-Time Interest Point (STIP) detector and its velocity adaptation, to human action recognition process. Using STIPs as descriptors and a Support Vector Machine classifier, we evaluate four different public video datasets to validate our methodology and demonstrate its accuracy in real scenarios.

Keywords
human action recognition, support vector machine, space-time interest points, C-STIP, V-STIP

1 INTRODUCTION

As computer vision trends come and go, there are topics that remain widely studied, such as human action recognition. This field is being studied over the last three decades, using numerous techniques and methods to solve a common problem.

In [GBS07], human action recognition is described as a key component in many computer vision applications: video surveillance, human-computer interface, video indexing and browsing, recognition of gestures, analysis of sports events and dance choreography.

Many approaches were developed in the last years, but most of them have computational limitations, some of wich are: difficulty to estimate the motion pattern [Bla99], aperture problems, discontinuities and smooth surfaces [lLL06]; all related with the motion estimation technique used, like optical flow or more complex techniques, such as eigenshapes of foreground silhouettes, described in [GKRR05].

Some approaches, in recent successful works, use human action recognition information from video sequences as a space-time volume of intensities, gradients, optical flow, or other local features [ZMI01a, SI05].

Our methodology is similar to the proposed approach of [SLC04] in which the classical space-time interest point detector is discussed and classified using a Support Vector Machine (SVM). However, only local spatio-temporal and histograms of local features were used. This limitation shows that for small videos dataset it properly works, but for large and complex video datasets, a global spatio-temporal descriptor is required.

In this paper, we use an existing state-of-the-art detector of human action recognition as a descriptor, without the assistance of other descriptors. Section 2 describes the main related works about human action recognition. Section 3 presents a detailed explanation of the space-time interest points techniques used in this work. In Section 4, the proposed methodology is shown and in Section 5, experimental results are discussed. Section 6 is dedicated to conclusions and further work.

2 RELATED WORKS

Commonly, human action recognition is divided into two major classes, model-based recognition and appearance-based recognition [BD01, Lap04].

Many works use a human model [Roh94, GD96], generally obtained by recreating the human body using a three-dimensional model with degrees-of-freedom that allows distinct poses, representing a certain movement that corresponds to an action. With more degrees-of-freedom, a larger number of body positions can be achieved, creating a greater number of different movements and, consequently, represented actions. A general model-based approach, as described in [Kan80], calls for a robust background and foreground segmentation to be able to distinguish between the picture domain and the scene domain. However, they are an easy way to estimate and predict the feature locations [BD01].

According to [BD01] appearance-based approaches are focused on representing an action as a motion over time and the motion recognition is achieved from appearance, since it has a space-time trajectory. In [DB97], an image template is used to recognize an action in video and it is obtained by accumulating motion images from specific key frames, in order that an image-vector is constructed and matched against previously generated ground-truth templates. This approach is used to construct a view-specific representation of action.

The approach described in [LL03] uses combined techniques based on appearance (e.g. [ZMI01b] and [MS04]) to achieve a novel motion event detector using local information and a 3D extension of Harris corner detector, named space-time interest points (STIP). One advantage of this representation is that it does not need previous segmentation or tracking [LAS08].

Over the years, numerous STIPs detectors have been proposed: the methodology shown in [DRCB05] uses a space-time cuboid to represent an interest point, achieved from the convolution of a quadrature pair of 1D Gabor filters with a 2D Gaussian smoothing kernel; an evolution of the first presented STIP is proposed by [LL04], adapting the velocity and spatio-temporal scale features, in order to obtain a stable video representation.

In [WTVG08] a Hessian-based approach is proposed, using the determinant of the Hessian as a saliency measure, being able to extract scale invariant features and densely cover the video content. And in in [CHMG12], a selective method which applies surround suppression combined with local and temporal constraints is shown, including a Bag of Videos model to build a vocabulary of visual-words for action recognition process.

3 SPACE-TIME INTEREST POINTS

According to [Lap04], the main purpose of the STIPs is to perform the event detection directly from the spatiotemporal data of the image, considering regions that have distinct locations in space-time with sufficient robustness to detect and classify. [LL03] use a 3D extension of Harris corner detector to detect these interest points.

[HS88] accentuate image areas that have maximum variation of image gradients in a local neighborhood. The goal of the Harris interest point detector is to find spatial locations where the image has significant changes in both directions.

Classical STIP

We consider the classical STIP (C-STIP) as the one proposed by [LL03] and its mathematical representation will be reviewed below.

Considering an image sequence $I(x,y,t)$ and its scale-space representation as

$$S(\cdot,\alpha^2,\beta^2) = I * G(\cdot,\alpha^2,\beta^2) \qquad (1)$$

where α^2 is the spatial variation, β^2 is the temporal variance and G is a Gaussian convolution kernel

$$G(x,y,t,\alpha^2,\beta^2) = \frac{exp(-(x^2+y^2)/2\alpha^2 - t^2/2\beta^2)}{\sqrt{(2\pi)^3\alpha^4\beta^2}} \qquad (2)$$

To detect interest points in I, it is necessary to search for meaningful eigenvalues $\lambda_1,\lambda_2,\lambda_3$ of a second-moment matrix γ

$$\gamma(\cdot,\alpha^2,\beta^2) = G(\cdot,s\alpha^2,s\beta^2) * (\nabla I(\nabla I)^T) \qquad (3)$$

that uses spatio-temporal image gradients $\nabla I = (I_x,I_y,I_t)^T$ within a Gaussian neighborhood of every point.

And compute the local maxima of the extended Harris corner function R

$$\begin{aligned} R &= det(\gamma) - k\,\text{trace}^3(\gamma) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1\lambda_2\lambda_3)^3 \end{aligned} \qquad (4)$$

where k is the sensitivity factor.

Thus, STIPs of I are discovered by detecting local positive spatio-temporal maxima in R.

Velocity adapted STIP

[LL04] adapted the local velocity and scales of C-STIP to compensate the relative motion between the object and the camera. To adapt the scales, a normalized Laplacian operator is calculated for each detected interest point in R (Equation 4)

$$\nabla^2 I = I_{x,norm} + I_{y,norm} + I_{t,norm} \qquad (5)$$

where $I_{x,norm}, I_{y,norm}$ and $I_{t,norm}$ are the second-order derivatives of I normalized by the scale parameters. The space-time maxima of the Harris corner function and a selection of points that maximizes the Laplacian normalizes operator is a method to adapt the previous C-STIP implementation [Lap05]. Lastly, to perform the velocity adaptation, we have to consider the Galilean transformations that affects the time domain. To cancel these effects we have to redefine the an operator of interest γ'' in terms of a velocity-adjusted descriptor, using a structure similar to that presented by [LK81]. The Equation 3 can now be redefined as

$$R_v = det(\gamma'') - k\,\text{trace}^3(\gamma''). \tag{6}$$

4 MATERIALS AND METHODS

As stated earlier, our approach is based on [LL03] technique, here nominated as C-STIP, and on [LL04], named V-STIP.

The state-of-the-art STIP is used as a local detector. To use it in a recognition and classification process, an additional descriptor, such as Histogram of Oriented Gradients (HOG) [DT05] or optical flow [BFB94], is used to improve the local features. Our methodology uses STIP as a descriptor without the aid of any other descriptors.

Following the flowchart described in Figure 1, each step of our proposed method will be more detailed.

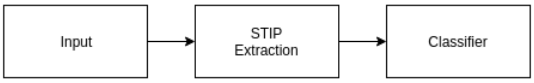

Figure 1: Our proposed methodology flowchart.

Input

For a complete evaluation of our proposed methodology, A number of specialized public video data sets were used to recognize human action. The main datasets used were:

- KTH [SLC04];
- UCF101 [SZS12];
- Weizmann [BGS05];
- YouTube [LLS09].

For each dataset, four different actions were studied in our proposed approach. An image sample of the selected actions and datasets is described in Table 4.

In the UCF101 dataset the following classes were used: *biking*, *jumping jack*, *punch* and *walking with dog*, with ten videos per action lasting about five seconds. In KTH the used classes were: *boxing*, *handwaving*, *running* and *walking*, having ten videos per action with an estimated duration of twenty seconds. In Weizmann the used classes were: *bend*, *gallop sideways*, *jump in place* and *skip*, with nine videos per action and lasting about two seconds. Finally, YouTube used the classes: *basketball*, *diving*, *soccer juggling* and *volleyball spiking*, consisting of seven videos per action with an approximate duration of seven seconds.

All selected video classes in the used datasets were randomly separated into training set (70%) and validation set (30%) and the actions were chosen in a way that it is possible to make a fair comparison between the both techniques C-STIP and V-STIP.

STIP Extraction

Since the state-of-the-art approach is C-STIP and V-STIP is its upgrade, our approach uses the methods showed in Section 3 to extract STIPs.

For the C-STIP implementation, the Equations 1 to 4 were used, generating a feature vector that is mathematically described in Section 3. Additionally, the reference structure can be found in [LL03].

The V-STIP implementation uses Equations 1 to 5, also producing a feature vector. For more details of the technique, we refer to [LL04].

Support Vector Machine

Support Vector Machines (SVMs) are state-of-the-art classifiers that produces an hyperplane based on the training data [CV95]. This achieved hyperplane has a prediction of the class labels of the validation data without further information besides its features. In [CV95] a more detailed structure of SVM is described.

Considering a training set of $(x_i, y_i), i = 1, ..., l$, where $x_i \in \Re$ is a feature vector and $y_i \in \{1,2,3,4\}$ its class labels, a Support Vector Classification (SVC) algorithm [BHHSV01] with a radial basis kernel, as shown in Equation 7, is used to train and predict the validation set class labels.

$$K(x,y) = exp(-\gamma\|x-y\|^2) \tag{7}$$

where here γ is the kernel parameter.

From the trained and validated C-STIP and V-STIP models, we evaluated all classes using a Mean Average Precision (MAP) measure, similar to [CHMG12] and [DOS15].

5 RESULTS

To train the SVM classifier, parameters were varied to evaluate which one achieve the better performance, local and global respectively. The selected parameters were: Harris corner function sensitivity factor (k), spatial variation (α) and temporal variance (β), all variables described in the Equation 4.

To evaluate our methodology, four different training scenarios (TS) for each implementation will be presented as it follows: TS1: $\alpha = 4$, $\beta = 2$ and $k = 0.01$;

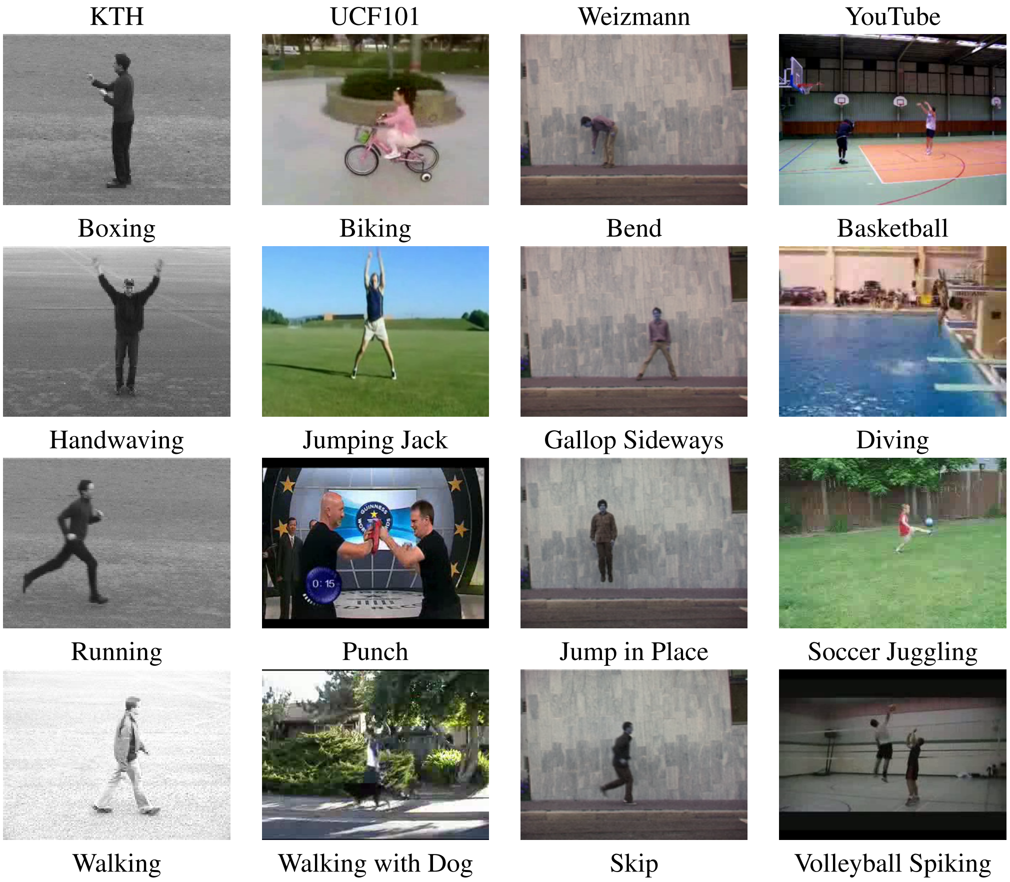

Table 1: Datasets samples.

TS2: $\alpha = 4$, $\beta = 2$ and $k = 0.05$; TS3: $\alpha = 9$, $\beta = 3$ and $k = 0.05$; TS4: $\alpha = 16$, $\beta = 4$ and $k = 0.05$.

After exhaustive evaluations, these training scenarios described above were selected from the best achieved results, improving our proposed methodology performance and denoting the direct influence of them on the MAP measure in action recognition process.

The achieved results are compiled, for each dataset and selected action, in confusion matrices found in Tables 5 to 5, where the best precision measure was highlighted.

In Tables 5 to 5, it is shown a MAP measure achieved from all classes based on the confusion matrices, summarizing the global results, presenting the results for C-STIP and V-STIP methodology, classified using the SVM schema described in Subsection 4, for each dataset.

Discussions

Due to the great amount of information that can be detailed using the confusion matrices, we opted to discuss the most relevant to evaluate the performance of the whole classification process given the proposed training scenarios.

For the first training scenario in Table 5, it is possible to notice that the process of classification of the dataset KTH, even with few false positives, was efficient to the majority of the videos in each class. This result was expected, since the KTH dataset is simpler and the human action is performed in clean environments with high contrast between the subject and the background, for all actions.

Using the C-STIP method, the classification for the actions *gallop* and *bend* was often incorrect. However, the V-STIP method has shown more suitable video classifications in this two classes, as a result of similar STIP variations in the lower body movements, generating a larger number of salience in this region. These variations occur in other scenarios with the same Weizmann dataset. In the training scenario 4, Table 5, there was a confusion between *handwave* and *box* due to lack of lateral variation by the subject, showing the influence of the spatial variation parameter (α) in classification results.

In the confusion matrices presented in Tables 5 to 5, it is possible to notice a recurrent confusion between the classes *punch* and *jumping jack* that can be explained by the upper body movement that is slightly alike between the selected videos. The V-STIP method was able to classify better than the C-STIP method, with the majority of correct classifications.

Method		C-STIP				V-STIP			
Dataset	Class	*box*	*handwave*	*run*	*walk*	*box*	*handwave*	*run*	*walk*
KTH	*box*	**42**	34	8	16	**50**	34.5	5	10.5
	handwave	30	**39**	16	15	34	**42**	13	11
	run	19	22	**47**	12	20	22	**51**	7
	walk	24	26	10	**40**	28	28	7	**37**
UCF101		*bike*	*jumpjack*	*punch*	*walkdog*	*bike*	*jumpjack*	*punch*	*walkdog*
	bike	**50**	16	21	13	**45**	31	12	12
	jumpjack	15	33	**35**	17	26	**47**	14	13
	punch	11	30	**43**	16	30	9	**44**	17
	walkdog	21	16	29	**34**	25	7	20	**48**
Weizmann		*bend*	*gallop*	*pjump*	*skip*	*bend*	*gallop*	*pjump*	*skip*
	bend	35	**35.5**	17.5	12	**52**	30	11	7
	gallop	20	**43**	20	17	29	**45.5**	20	5.5
	pjump	13	23	**37**	28	16	27	**53**	4
	skip	12	18	29	**41**	**36**	21	9	34
YouTube		*basket*	*dive*	*soccer*	*volleyb*	*basket*	*dive*	*soccer*	*volleyb*
	basket	23	31	**43**	3	14	36	**47**	3
	dive	5	**51**	41	3	6.5	**49**	43	2
	soccer	6	36	**54**	4	7	38	**52**	3
	volleyb	1	7	16	**76**	1	4	14	**81**

Table 2: Confusion matrix for TS1.

Method		C-STIP				V-STIP			
Dataset	Class	*box*	*handwave*	*run*	*walk*	*box*	*handwave*	*run*	*walk*
KTH	*box*	**43**	35	8.5	13.5	**52**	36	4	8
	handwave	26. 5	**34**	27.5	12	29	**42**	22	7
	run	16	19	**56**	9	21.5	23	**51**	5
	walk	24	27.5	7.5	**41**	30	31	4	**35**
UCF101		*bike*	*jumpjack*	*punch*	*walkdog*	*bike*	*jumpjack*	*punch*	*walkdog*
	bike	**52**	14	18	16	**49**	12	19	20
	jumpjack	14	**39**	33	14	15	33	**37**	15
	punch	10	37	**39**	14	11	29	**44**	16
	walkdog	14	17	32	**37**	13.5	14	30.5	**42**
Weizmann		*bend*	*gallop*	*pjump*	*skip*	*bend*	*gallop*	*pjump*	*skip*
	bend	23	**37**	24	16	**42**	34	13.5	10.5
	gallop	14	**34**	29	23	26	**34**	21	19
	pjump	10	22	**38**	30	14.5	20	**35.5**	30
	skip	9	19	33	**39**	16	17	29	**38**
YouTube		*basket*	*dive*	*soccer*	*volleyb*	*basket*	*dive*	*soccer*	*volleyb*
	basket	**44**	22	32	2	22	34	**42**	2
	dive	8	**49**	41	2	9	**52**	37	2
	soccer	9	36	**52**	3	12	38	**48**	2
	volleyb	2	5	18	**75**	2	3	16	**79**

Table 3: Confusion matrix for TS2.

Method		C-STIP				V-STIP			
Dataset	Class	*box*	*handwave*	*run*	*walk*	*box*	*handwave*	*run*	*walk*
KTH	*box*	**48**	38	5	9	**52**	38	3	6
	handwave	32	**46**	14	8	37	**43**	14	5
	run	18	23	**52**	7	22	23	**51**	4
	walk	26	**31**	12	30	31	28	8	**34**
UCF101		*bike*	*jumpjack*	*punch*	*walkdog*	*bike*	*jumpjack*	*punch*	*walkdog*
	bike	**65**	9	15	11	**57**	10	20	13
	jumpjack	10	34	**43**	14	16	27	**42**	15
	punch	6	31	**49**	14	8	26	**53**	14
	walkdog	12	17	29	**43**	10	14	27	**49**
Weizmann		*bend*	*gallop*	*pjump*	*skip*	*bend*	*gallop*	*pjump*	*skip*
	bend	**35**	34	17	14	**62**	22	8	9
	gallop	24	**49**	16	11	**44**	35	12	9
	pjump	14	26	**34**	26	24	23	**28**	24
	skip	10	20	31	**39**	23	16	25	**37**
YouTube		*basket*	*dive*	*soccer*	*volleyb*	*basket*	*dive*	*soccer*	*volleyb*
	basket	**63**	13	23	1	20	33	**45**	2
	dive	8	**51**	39	2	9	**53**	36	2
	soccer	11	35	**52**	2	13	37	**49**	2
	volleyb	1	5	13	**81**	2	4	12	**82**

Table 4: Confusion matrix for TS3.

Method		C-STIP				V-STIP			
Dataset	Class	*box*	*handwave*	*run*	*walk*	*box*	*handwave*	*run*	*walk*
KTH	*box*	**50**	37	5	8	**62**	27	4	7
	handwave	41	**47**	4	8	**48**	44	3	5
	run	15	20	**61**	4	29	20	**46**	5
	walk	29	**31**	10	30	35	22	7	**37**
UCF101		*bike*	*jumpjack*	*punch*	*walkdog*	*bike*	*jumpjack*	*punch*	*walkdog*
	bike	**62**	11	16	11	**53**	14	21	12
	jumpjack	15	30	**41**	13	19	29	**38**	14
	punch	13	27	**47**	13	13	22	**51**	14
	walkdog	17	17	31	**35**	14	15	32	**40**
Weizmann		*bend*	*gallop*	*pjump*	*skip*	*bend*	*gallop*	*pjump*	*skip*
	bend	**48**	35	10	7	**57**	31	6	6
	gallop	27	**40**	18	14	**42**	33	13	12
	pjump	15	20	**35**	31	28	18	**29**	25
	skip	12	16	28	**44**	14	15	32	**39**
YouTube		*basket*	*dive*	*soccer*	*volleyb*	*basket*	*dive*	*soccer*	*volleyb*
	basket	**60**	16	23	1	29	31	**39**	1
	dive	11	**47**	41	1	14	**46**	39	1
	soccer	14	34	**50**	2	16	35	**46**	2
	volleyb	3	6	9	**82**	3	6	8	**83**

Table 5: Confusion matrix for TS4.

Method	KTH	UCF101	Weizmann	YouTube
C-STIP	42%	40%	39%	51%
V-STIP	45%	46%	46%	49%

Table 6: TS1 - Mean Average Precision (MAP).

Method	KTH	UCF101	Weizmann	YouTube
C-STIP	42%	44%	33%	53%
V-STIP	54%	39%	37%	48%

Table 7: TS2 - Mean Average Precision (MAP).

Method	KTH	UCF101	Weizmann	YouTube
C-STIP	44%	47%	39%	61%
V-STIP	45%	46%	40%	51%

Table 8: TS3 - Mean Average Precision (MAP).

Method	KTH	UCF101	Weizmann	YouTube
C-STIP	47%	43%	42%	60%
V-STIP	47%	43%	39%	51%

Table 9: TS4 - Mean Average Precision (MAP).

All the datasets had videos with different duration, as presented in Subsection 4, and the proposed methodology was able to correctly classify the videos, even with the variable time, showing the robustness of the technique.

Using the MAP results in Tables 5 to 5 and the C-STIP method as first reference, the best results for KTH and Weizmann datasets were in scenario 4, with $\alpha = 16$, $\beta = 4$ and $k = 0.05$. For UCF101 and YouTube, scenario 3 have better performances, with $\alpha = 9$, $\beta = 3$ and $k = 0.05$. These were the best parameters adjustments for the aforementioned datasets.

With V-STIP as reference, the best parameter adjustment for KTH dataset is $\alpha = 16$, $\beta = 4$ and $k = 0.05$, that represents scenario 4. UCF101 and Weizmann best results were with scenario 1, where the parameters were $\alpha = 4$, $\beta = 2$ and $k = 0.01$. UCF101 can also be fit with $\alpha = 9$, $\beta = 3$ and $k = 0.05$. For YouTube dataset, there is a tie between two scenarios, TS3 and TS4, therefore it is possible to use $\alpha = 9$, $\beta = 3$ and $k = 0.05$ or $\alpha = 16$, $\beta = 4$ and $k = 0.05$.

6 CONCLUSIONS

Our goal is to evaluate state-of-the-art techniques of space-time interest points descriptors using a SVM classifier to recognize human actions in videos. Two descriptor implementations were proposed and applied in known datasets to classify actions, not taking into consideration the videos resolution and duration.

The proposal of using STIPs as descriptors in its two variations (Classic and Velocity adaptation) is valid, since the classification step was able to correctly classify the majority of the presented actions or classes, indicating that this work can be used as an alternative method to address the problem of human action recognition in videos.

Considering unclipped videos, that are videos with different actions occurring at the same time, it is possible as well to claim that the proposal can also classify the main action of the scene, once two tested datasets have, somewhat, this characteristic (UCF101 and YouTube).

The results achieved can be also be used as parameters directives for an optimization of the adjustment stage of future works.

Finally, further works will prioritize the diagonals of confusion matrices, avoiding false positives and improving global results.

7 REFERENCES

[BD01] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, Mar 2001.

[BFB94] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *Int. J. Comput. Vision*, 12(1):43–77, February 1994.

[BGS05] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[BHHSV01] Asa Ben-Hur, David Horn, Hava T Siegelmann, and Vladimir Vapnik. Support vector clustering. *Journal of machine learning research*, 2(Dec):125–137, 2001.

[Bla99] M. J. Black. Explaining optical flow events with parameterized spatio-temporal models. In *IEEE Proc. Computer Vision and Pattern Recognition, CVPR'99*, pages 326–332, Fort Collins, CO, 1999. IEEE.

[CHMG12] Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund, and Jordi GonzÃ lez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396 – 410, 2012. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[DB97] James Davis and Aaron Bobick. The representation and recognition of action using temporal templates. pages 928–934, 1997.

[DOS15] Afshin Dehghan, Omar Oreifej, and Mubarak Shah. Complex event recognition using constrained low-rank representation. *Image and Vision Computing*, 42:13–21, 2015.

[DRCB05] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, ICCCN '05, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[GBS07] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, Dec 2007.

[GD96] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, pages 73–80, Jun 1996.

[GKRR05] Roman Goldenberg, Ron Kimmel, Ehud Rivlin, and Michael Rudzsky. Behavior classification by eigendecomposition of periodic motions. *Pattern Recogn.*, 38(7):1033–1043, July 2005.

[HS88] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Citeseer, 1988.

[Kan80] Takeo Kanade. Region segmentation: Signal vs semantics. *Computer Graphics and Image Processing*, 13(4):279 – 297, 1980.

[Lap04] Ivan Laptev. Local spatio-temporal image features for motion interpretation. 2004.

[Lap05] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[LAS08] Jingen Liu, Saad Ali, and Mubarak Shah. Recognizing human actions using multiple features. 2008.

[LK81] An iterative image registration technique with an application to stereo vision. Lucas, Bruce D., and Takeo Kanade. 1981.

[LL03] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *IN ICCV*, pages 432–439, 2003.

[LL04] Ivan Laptev and Tony Lindeberg. Velocity adaptation of space-time interest points. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 52–56. IEEE, 2004.

[lLL06] Wei lwun Lu and James J. Little. Tracking and recognizing actions at a distance. In *in: ECCV Workshop on Computer Vision Based Analysis in Sport Environments*, pages 49–60, 2006.

[LLS09] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild". In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE, 2009.

[MS04] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[Roh94] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Underst.*, 59(1):94–115, January 1994.

[SI05] Eli Shechtman and Michal Irani. Space-time behavior based correlation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 405–412, June 2005.

[SLC04] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. 2012.

[WTVG08] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on computer vision*, pages 650–663. Springer, 2008.

[ZMI01a] L. Zelnik-Manor and M. Irani. Event-based video analysis. Technical report, Jerusalem, Israel, Israel, 2001.

[ZMI01b] Lihi Zelnik-Manor and Michal Irani. Event-based analysis of video. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–123. IEEE, 2001.