Human Action Recognition using Lagrangian Descriptors

Esra Acar [#1], Tobias Senst [*1], Alexander Kuhn [+1], Ivo Keller [*2],
Holger Theisel [+2], Sahin Albayrak [#2], Thomas Sikora [*3]

[#] *DAI Laboratory, Technische Universität Berlin*
Ernst-Reuter-Platz 7, TEL 14, 10587 Berlin, Germany
[1] esra.acar@dai-labor.de [2] sahin.albayrak@dai-labor.de

[*] *Communication Systems Group, Technische Universität Berlin*
EN 1, Einsteinufer 17, 10587 Berlin, Germany
[1] senst@nue.tu-berlin.de [2] keller@nue.tu-berlin.de [3] sikora@nue.tu-berlin.de

[+] *Department of Simulation and Graphics,University of Magdeburg*
Universitätsplatz 2, 39106 Magdeburg, Germany
[1] akuhn@isg.cs.uni-magdeburg.de [2] theisel@isg.cs.uni-magdeburg.de

***Abstract*—Human action recognition requires the description of complex motion patterns in image sequences. In general, these patterns span varying temporal scales. In this context, Lagrangian methods have proven to be valuable for crowd analysis tasks such as crowd segmentation. In this paper, we show that, besides their potential in describing large scale motion patterns, Lagrangian methods are also well suited to model complex individual human activities over variable time intervals. We use Finite Time Lyapunov Exponents and time-normalized arc length measures in a linear SVM classification scheme. We evaluated our method on the Weizmann and KTH datasets. The results demonstrate that our approach is promising and that human action recognition performance is improved by fusing Lagrangian measures.**

I. Introduction

Human action recognition is an important computer vision task, since it pertains to a multitude of application domains, ranging from automatic video surveillance to semantic video indexing and retrieval. In this paper, we aim at providing a reliable solution for recognizing basic human actions including (among others) walking, running and boxing. The ability to identify these basic actions would subsequently lead us to build a solution to recognize more complex actions, in particular in the scenarios involving more than one basic action.

One category of popular approaches to capture the motion dynamics in image sequences are based on optical flow computation. Optical flow methods describe the motion in an image sequence by means of vector fields, which encode the transport and correlation of image information between two consecutive frames. Recent improvements in optical flow-based approaches allow for an efficient extraction of optical flow fields with a diversity of methods, offering a variety of trade-offs between speed and accuracy of the motion extraction. Extracting the transport vector fields for a complete image sequence further helps in analyzing global motion features in a spatio-temporal context i.e. in the *space time domain*. The sequence of optical flow fields can be treated as an unsteady vector field and allows for the application of the Lagrangian analysis framework, which has been introduced in the context of dynamical systems. Lagrangian analysis is based on particle trajectories in the space time domain. In this context, the concept of Lagrangian Coherent Structures (LCS) has proven to be a powerful tool to describe temporally complex spatial motion patterns. An overview of this research topic has been provided by Peacock et al. [1] and a general presentation of methods tailored towards flow analysis has been presented by Pobitzer et al. [2]. One popular manner to describe the notion of LCS in time-varying vector fields is the Finite Time Lyapunov Exponents (FTLE) method introduced by Haller et al. [3]. In addition to these promising approaches, our work focuses on the description of individual motion behaviors using FTLE.

The classic FTLE approach has been developed in the context of fluid flow analysis. One major difference is that, optical flow fields are in general not divergence free, and only represent a *projection* of a higher dimensional motion process. Still most of the captured Lagrangian features are suitable to describe motion processes in images, and can be considered as strongly correlated to the movement, i.e. human motion in the underlying image sequence. In our work, we focus on two important Lagrangian motion features for the description of human actions in images:

- **Motion boundaries** denote areas of high particle trajectory separation and segment areas of different motion patterns of varying time intervals. Those are directly captured using the FTLE descriptor.
- **Areas of coherent flow motion** gather regions of similar speed over the respective time interval. This corresponds to a clustering of trajectories of similar flow geometry.

MMSP'12, September 17-19, 2012, Banff, Canada.

A detailed discussion of those features in the context of optical flow field analysis is provided in Section III.

II. Related Work

Video representation and learning are the two components of an action recognition system. Many learning-based approaches for human action recognition have been proposed in the literature. In this paper, we concentrate on the recognition of actions performed by a single actor in videos and discuss existing works addressing this task. The paper by Poppe et al. [4] provides a survey of vision-based human action recognition in broad situations.

Support Vector Machines (SVMs), as one of the existing discriminative approaches, have been extensively applied in this context, and have proven to be successful. For instance, Schindler and Van Gool [5] represent human actions by object shape and optical flow features, which they use for training *K* linear one-vs-all SVM action classifiers. Positive and negative training samples are weighted differently to account for the problem of imbalanced training data. Bregonzio et al. [6] represent actions as clouds of space-time interest points. They also use SVMs for classification, but, unlike [5] where linear kernels are used, they use radial basis function (RBF) kernels. Boosting, another discriminative approach, has also been employed in the context of action classification. Fathi and Mori [7] use the AdaBoost algorithm both for feature selection and action classification. Feature selection is performed on mid-level motion features which are constructed from low-level optical flow information using AdaBoost. AdaBoost is also used as the final action classifier based on the selected mid-level motion features.

Alternative to the discriminative approach, in the generative approach, the joint occurrence of visual data and action labels is modeled. Niebles et al. [8] represent video sequences as collections of spatio-temporal words which are based on spatio-temporal interest points. They use two topic modeling methods: the probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA). A separate topic model is learned for each action class and new samples are classified by using the constructed action topic models.

k-Nearest Neighbor (*k*-NN) classifiers are used for action recognition in [6] and [9]. In [6], video samples are represented as clouds of space-time interest points and *k*-NN classification is performed using these video representations. In [9], each image sequence is represented as a bag of kinematic modes and is mapped into a kinematic mode-based feature space where an action is given the same label as its nearest neighbor in the feature space.

III. Method

As described in Section I, we can treat a series of optical flow fields as a time-dependent vector field, and define the space time domain by interpreting time as an additional axis. Formally, this representation can be described as follows: Given the optical flow vector field $\mathbf{v}(\mathbf{x}, t)$, at every specified space-time point $(\mathbf{x}_0, t_0) \in D$ we can start a *path line* that denotes a particle trajectory. This can be formulated in terms of an initial value problem as follows:

$$\frac{d}{dt}\begin{pmatrix} \mathbf{x} \\ t \end{pmatrix} = \begin{pmatrix} \mathbf{v}(\mathbf{x}(t), t) \\ 1 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{x} \\ t \end{pmatrix}(0) = \begin{pmatrix} \mathbf{x}_0 \\ t_0 \end{pmatrix}$$

The *path lines* are extracted by computing the *flow map* defined as $\phi^\tau(\mathbf{x}, t_0) = \phi(\mathbf{x}, t_0, \tau)$. The flow map describes a mapping of an initial position to its advected position after a predefined integration time τ starting at t_0. Combining all positions on these trajectories of one specific point within the interval $[t_0, t_0 + \tau]$ creates a polynomial curve denoted as *path line*.

To analyze motion properties in a feature-oriented manner the concept of LCS has been proposed. LCS directly describes properties of neighboring trajectories in the space time domain [1]. In video analysis, this corresponds to the notion of the edge of an enclosed moving object within the image. One crucial aspect is the choice of the time interval parameter τ, that defines the temporal scale of features we are interested in. The most prominent techniques to extract LCS are FTLE presented by Haller et al. [3]. LCS are shown to be closely related to *ridges* in the resulting FTLE scalar field [10].

We propose to apply a set of Lagrangian measures to describe different human actions by LCS. These measures present several advantages. The most notable one is the ability of Lagrangian measures to transform the motion information about LCS of a given time interval τ into a 2D space. By applying a feature extraction method at this 2D space the resulting feature considers implicitly the motion information of the observed object. In addition, this resulting feature can be classified as a spatio-temporal feature.

A. Lagrangian Descriptors

Given the optical flow field $\mathbf{v}(\mathbf{x}, t)$ the first measure we use in our framework is the FTLE.

$$\text{FTLE}(\mathbf{x}, t, \tau) = \frac{1}{\tau} max\left\{\mu_1, \mu_1\right\} \tag{1}$$

with μ the eigenvalue defined as:

$$\mu_i = ln\sqrt{\lambda_i(\nabla^T \nabla)} \tag{2}$$

and ∇ the spatial gradient defined as:

$$\nabla(\mathbf{x}, t, \tau) = \frac{\partial \phi(\mathbf{x}, t, \tau)}{\partial \mathbf{x}} \tag{3}$$

The FTLE can be computed both in forward and backward directions resulting in the description of FTLE+ and FTLE- as described by Garth et al. [11] (see Figure 2(b,c)). Lagrangian features in the FTLE+ field define regions of repelling LCS, while FTLE- features describe attracting LCS structures over the considered time interval. Intersection points of the FTLE+ and FTLE- features group areas of coherent motion within the field. One recent work in this context is presented by Bachthaler et al. [12]. Using this notion, ridges can be reinterpreted as motion barriers defined over a finite time scope [3]. The parameter τ determines the length and complexity of

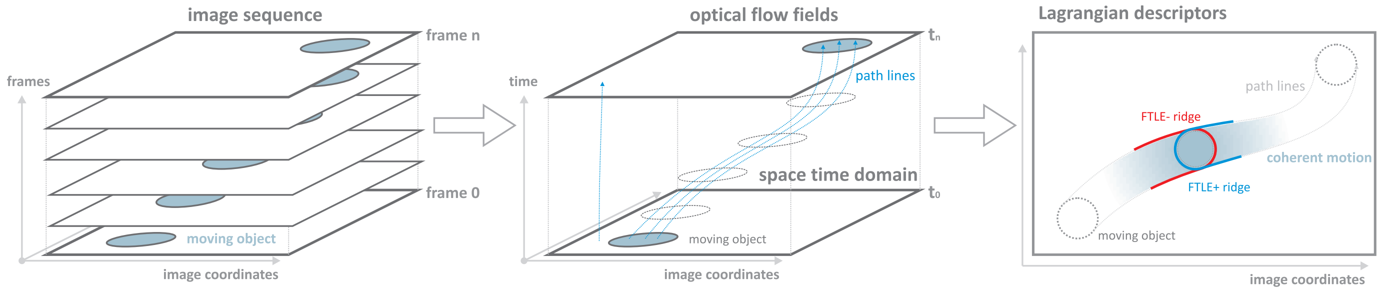

Fig. 1. Concept of Lagrangian descriptors obtained from a given image sequence using a series of optical flow fields.

those ridge structures. As shown in Figure 2, the FTLE field appears as an excellent tool to model the motion boundaries of a moving person as it describes the behavior of trajectories in terms of transport barriers and separates regions of coherent flow behavior. These motion boundaries are not only good cues for the outline of person detection and crowd description as stated in [13], [14], but they also constitute good cues for person description [15].

In general, features in the FTLE field are explicitly extracted in terms of height ridges that require an additional ridge extraction procedure. This ridge extraction tends to be noise-sensitive in the underlying field. We avoid explicitly extracting ridge structures by computing Histogram of oriented Gradients (HoG) [16] of the FTLE field. The HoG descriptor partitions a detector window into a dense grid of cells, with each cell containing a local histogram over orientation bins. The HOG is modified such that the FTLE field of the forward (FTLE+) and backward (FTLE-) integration is calculated at each pixel and converted to an angle (i.e. orientation). Each pixel associated to an angle contributes to the corresponding orientation bin with a vote weighted by the overall FTLE magnitude. The post processing grouping of the cells into blocks and the robust normalization are those of the conventional HOG. While FTLE describes the separation between neighboring trajectories, for the description of motion patterns we are usually also interested in areas of similar motion behaviors over time. This can be formulated using further geometric features of the path lines such as the time-normalized arc length, that corresponds to the accumulated average velocity at a certain point in the space time domain. A detailed discussion of alternative relevant path line attributes for flow feature description has been presented by Pobitzer et al. [17]. The time-normalized arc length (Λ_{arcL}) of given time t_0 and an integration interval τ is defined as

$$\Lambda_{arcL}(\mathbf{x}, t_0) = \int ||\mathbf{v}(\phi(\mathbf{x}, t_0, \tau))||_2 \partial\tau \tag{4}$$

The Λ_{arcL} denotes at each position the length of the path line which is equivalent to the overall speed at the respective position. Finally, the computation of HoGs is performed on the Λ_{arcL} field to extract the spatio temporal pattern.

B. *Lagrangian Classifier*

In order to compute the motion information, we use the GPU accelerated dense optical flow method proposed by Werlberger et al. [18]. For each video frame, we apply a multi-scale HoG detector as described in the seminal work of Dalal and Triggs where the HoG person detector was introduced [16]. The difference is that our detector uses the FTLE-HoG and the Λ_{arcL} -HoG instead of the conventional HoG.

Concerning the classification, we train linear multi-class SVMs. Training linear SVMs is faster and simpler than training SVM with other kernels e.g. RBF kernels. As demonstrated in [16], kernels other than linear would only lead to a slight performance improvement, but at the detriment of increased computational cost.

FTLE-HoG and Λ_{arcL} -HoG descriptors provide complementary information about an ongoing action in a video. Therefore, these two descriptors are fused in an early fusion manner before training SVMs. We choose to perform fusion with an equal weighting of FTLE-HoG and Λ_{arcL} -HoG descriptors, since we think that these descriptors provide equally valuable motion information. Additionally, assigning different weights to the FTLE-HoG and Λ_{arcL} -HoG descriptors would probably cause a bias towards the datasets used to evaluate the accuracy of the method.

IV. Privacy Aspect

Working with image data obtained from surveillance videos or publicly available image data further introduces another important aspect during automated processing of the image sequences: The privacy of the recorded individuals. In many existing applications, image data and subsamples of existing images has to be explicitly stored or processed. Due to the additional layer of abstraction in our processing cycle, we do not have to store image sequences directly, but optical flow fields, that only encode abstract motion patterns. In general, the recognition of person-related features extracted from abstract Lagrangian motion patterns is much more difficult if not even impossible.

V. Performance Evaluation

We tested our method on two standard datasets. These are the Weizmann and the KTH datasets. Experiments show that the combined use of FTLE-HoG and Λ_{arcL} -HoG features for

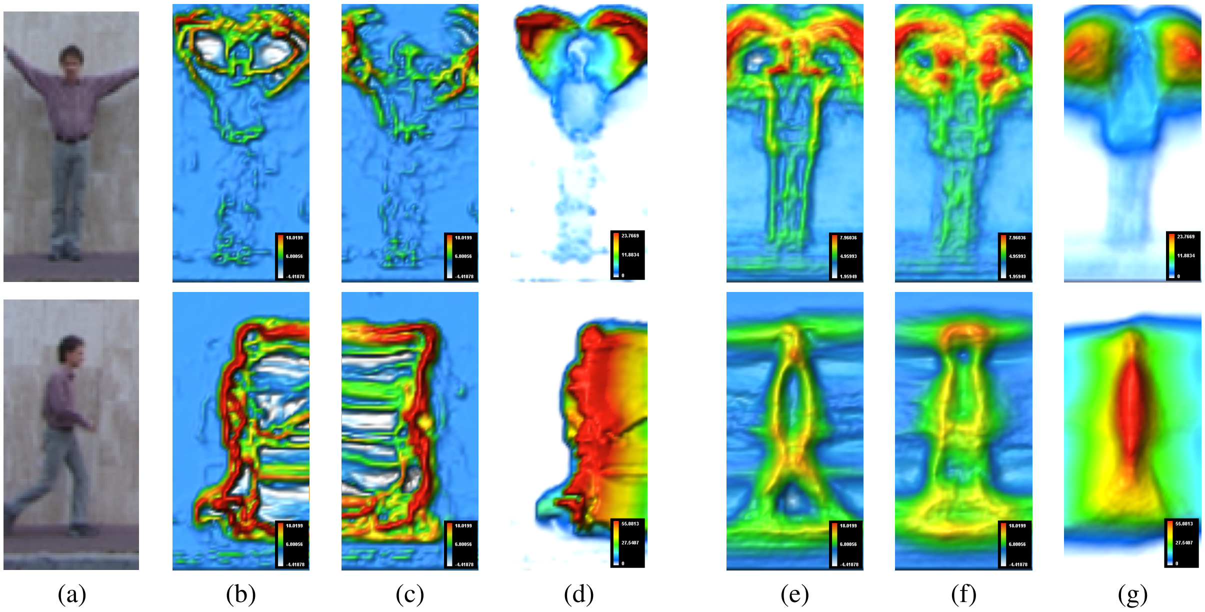

Fig. 2. Illustration of the Lagrangian descriptor for the Weizmann sequence.(a) Reference image of frame 1, sequence wave2 (top) and walk (bottom) with the corresponding FTLE+(b), FTLE-(c) and Λ_{arcL} (d) field. (e-g) Average of the FTLE+,FTLE- and Λ_{arcL} field for all corresponding sequences.

action recognition is promising, as results are comparable to state-of-the-art action recognition solutions.

A. Datasets

Weizmann Dataset - The Weizmann dataset was introduced in [19]. This dataset contains 9 actors performing 9 different types of actions (bending, galloping sideways, jumping jack, jumping forward on two legs, jumping in place on two legs, skipping, walking, waving one hand and waving two hands). Each video clip, which lasts about 2 seconds at 25 frames per second (fps), contains one actor performing one action.

KTH Dataset - The KTH dataset was introduced in [20]. The dataset contains 25 actors performing 6 different types of actions (boxing, hand clapping, hand waving, jogging, running and walking). The video clips are also recorded at 25 fps with varying durations. Like in the Weizmann dataset, each video clip contains one actor performing one action in 4 different scenarios including outdoors, outdoors with scale variation, outdoors with different clothes and indoors.

B. Experimental Setup

SVMs were trained using video frames which are represented with FTLE-HoG and Λ_{arcL} -HoG descriptors. Our approach was evaluated on the Weizmann dataset using Leave-One-Actor-Out Cross-Validation (LOAOCV). We used the video clips of 8 actors in the Weizmann dataset as the training data and the video clips of the remaining actor as the test data. This procedure was repeated by permuting the actors selected for training and testing, and the results were averaged. For the KTH dataset, video clips of 25 actors were splitted into training, validation and test parts using the provided standard split [21].

FTLE-HoG and Λ_{arcL} -HoG features of video frames are extracted as explained in Section III using cell grids of sizes 8 and 16.

C. Results and Discussion

Table I reports the classification accuracies of our method compared to state-of-the-art methods on the Weizmann dataset. We achieved 96.03% recognition accuracy with the fused FTLE-HoG and Λ_{arcL} -HoG features of cell size 8, and 97.55% recognition accuracy with a cell size of 16. This demonstrates the potential of our approach for human action recognition, as the results are comparable to those of state-of-the-art approaches.

TABLE I
CLASSIFICATION ACCURACIES ON THE WEIZMANN DATASET

Method	Accuracy (%)
Our method (cell size 8)	**96.03**
Our method (cell size 16)	**97.55**
Schindler et al. [5]	100.0
Bregonzio et al. [6]	96.66
Fathi et al. [7]	100.0
Niebles et al. [8]	90.0
Ali et al. [9]	94.75

In Figure 3, the confusion matrix of recognition results for the Weizmann dataset is illustrated. The confusion matrix represents the performance of our method with the fused FTLE-HoG and Λ_{arcL} -HoG features of cell size 16. As illustrated, skip and jump actions are the most difficult actions to discriminate. FTLE-HoG and Λ_{arcL} -HoG features of these two actions are similar, which causes the method to perform poorer compared to other actions. By including more discriminative motion information such as Λ_{arcL} -HoG features in both X and Y directions, we expect to better represent human actions.

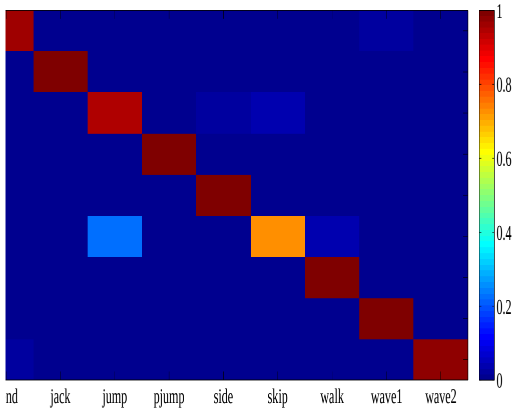

Fig. 3. Confusion matrix on the Weizmann dataset with FTLE-HoG + Λ_{arcL} -HoG features of cell size 16 (Mean accuracy: 97.55).

We performed additional experiments on the Weizmann dataset to show the effect of feature fusion in action recognition. As shown in Table II, fusing FTLE-HoG and Λ_{arcL} -HoG features of video frames improved the recognition accuracy of our method on the Weizmann dataset. This is due to the complementary information that FTLE-HoG and Λ_{arcL} -HoG features provide for an ongoing human action in a video sequence.

TABLE II
FEATURE FUSION ANALYSIS ON THE WEIZMANN DATASET

Cell Size	**FTLE**	Λ_{arcL}	**FTLE +** Λ_{arcL}
8	94.27	93.75	96.03
16	95.1	95.51	97.55

We also performed tests on the KTH dataset which is a challenging dataset due to indoor and outdoor video sequences with different lighting conditions. Additionally, the video sequences contain zoom-in and zoom-out during the performance of actors. The results of evaluation are presented in Table III which also shows results of state-of-the-art methods on the KTH dataset. There is no unique test methodology on this dataset, unlike the Weizmann dataset. Fathi et al. [7] use 16 actors of the dataset as the training data and the remaining 9 actors as the test data. Bregonzio et al. [6] and Niebles et al. [8] apply LOAOCV method, where Schindler et al. [5] use 5-fold cross validation. We followed the test strategy proposed in [9] and used 8 actors for training, 8 actors for validation and the remaining 9 actors for testing. On the dataset, we achieved 87.84% recognition accuracy with the fused FTLE-HoG and Λ_{arcL} -HoG features of cell size 8, and 86.11% recognition accuracy with a cell size of 16. The performance is lower than the performance on the Weizmann dataset, but is still comparable to those of state-of-the-art. This shows that our method is able to deal with challenging video sequences. However, the evaluation on the KTH dataset has to be interpreted carefully. Although we achieved a performance slightly lower than the method achieving the best performance [6], it is difficult to affirm that a given method performs better. Indeed, each work evaluated on this dataset used a different training and test methodology which makes a direct comparison irrelevant. Additionally, the difference and also the advantage of our method resides in transforming the motion information of an individual in a given time interval into a 2D space. Therefore, feature analysis is performed on a simplified representation instead of working on a 3D space.

TABLE III
CLASSIFICATION ACCURACIES ON THE KTH DATASET

Method	**Test Methodology**	**Accuracy (%)**
Our method (cell size 8)	Splits (into 3)	**87.84**
Our method (cell size 16)	Splits (into 3)	**86.11**
Ali et al. [9]	Splits (into 3)	87.7
Fathi et al. [7]	Splits (into 2)	90.50
Schindler et al. [5]	5-fold CV	92.70
Bregonzio et al. [6]	LOAOCV	94.33
Niebles et al. [8]	LOAOCV	83.33

The confusion matrix in Figure 4 allows to visualize the performance of our method with the fused FTLE-HoG and Λ_{arcL} -HoG features of cell size 8 on the KTH dataset. As illustrated, jogging - running, jogging - walking, hand clapping - hand waving, and boxing - hand clapping action pairs are the most confused action pairs. This is intuitive, since these confused action pairs have similar motion patterns. The results on the KTH dataset showed that further motion features such as Λ_{arcL} features both in X and Y directions are needed for better discrimination between these similar actions. For instance, the action "hand waving" shows motion both in X and Y directions, but the action "hand clapping" shows motion mainly in the X direction. Motion features in the X direction of these two actions are similar to each other and this causes the method to perform poorly.

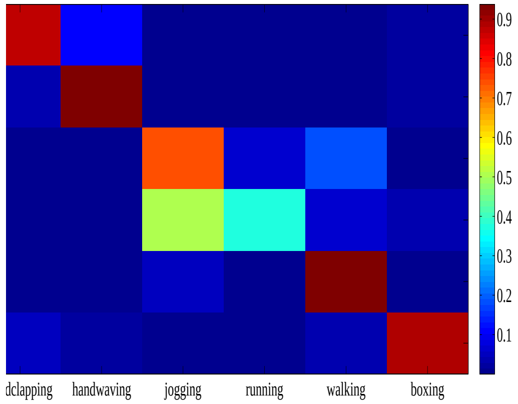

Fig. 4. Confusion matrix on the KTH dataset with FTLE-HoG + Λ_{arcL} -HoG features of cell size 8 (Mean accuracy: 87.84).

We performed feature fusion analysis also on the KTH dataset. As shown in Table IV, fusing FTLE-HoG and Λ_{arcL}

-HoG features of video frames improved the recognition accuracy of our method also on the KTH dataset. This confirms our statement that the information provided by the FTLE-HoG and Λ_{arcL} -HoG features complement each other.

TABLE IV
FEATURE FUSION ANALYSIS ON THE KTH DATASET

Cell Size	**FTLE**	Λ_{arcL}	**FTLE +** Λ_{arcL}
8	82.52	82.65	87.84
16	79.62	81.92	86.11

VI. CONCLUSIONS AND FUTURE WORK

We presented a framework based on Lagrangian methods which makes use of FTLE and time-normalized arc length measures for human action recognition. Our method differs from other human action recognition approaches, since with Lagrangian measures we transform the motion information of an individual in a given time interval into a 2D space. Our experiments on the Weizmann and KTH datasets proved that FTLE and time-normalized arc length measures are well adapted to model individual human activities. The experiments also proved that these features provide complementary information about an ongoing action in a video sequence and that by fusing these features very promising results can be achieved. Our ongoing work includes enriching the representation of motion within a video with time-normalized arc length measures both in X and Y directions. By including such information, we expect an improvement in performance.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's FP7 under grant agreement number 261743 (NoE VideoSense). The project SemSeg acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number 226042.

REFERENCES

[1] T. Peacock and J. Dabiri, "Introduction to focus issue: Lagrangian coherent structures," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 1, p. 017501, 2010.

[2] A. Pobitzer, R. Peikert, R. Fuchs, B. Schindler, A. Kuhn, H. Theisel, K. Matkovic, and H. Hauser, "The state of the art in topology-based visualization of unsteadyflow," *Computer Graphics Forum*, vol. 30, no. 6, pp. 1789–1811, 2011.

[3] G. Haller, "Lagrangian structures and the rate of strain in a partition of two-dimensional turbulence," *Physics of Fluids*, vol. 13, no. 11, 2001.

[4] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976 – 990, 2010.

[5] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?" in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1 –8.

[6] M. Bregonzio, T. Xiang, and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," *Pattern Recogn.*, vol. 45, no. 3, pp. 1220–1234, Mar. 2012.

[7] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1 –8.

[8] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, Sep. 2008.

[9] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 288 –303, feb. 2010.

[10] G. Haller, "A variational theory of hyperbolic Lagrangian Coherent Structures," *Physica D*, vol. 240, pp. 574–598, 2010.

[11] C. Garth, G. Li, X. Tricoche, C. Hansen, and H. Hagen, "Visualization of coherent structures in transient 2d flows," *Topology-Based Methods in Visualization II*, pp. 1–13, 2009.

[12] S. Bachthaler, F. Sadlo, C. Dachsbacher, and D. Weiskopf, "Space-time visualization of dynamics in lagrangian coherent structures of time-dependent 2d vector fields," *International Conference on Information Visualization Theory and Applications*, pp. 573–583, 2012.

[13] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision (ECCV 2006)*, 2006, pp. 428–441.

[14] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –6.

[15] T. Senst, R. Heras Evangelio, and T. Sikora, "Detecting people carrying objects based on an optical flow motion model," in *IEEE Workshop on Applications of Computer Vision (WACV 11)*, 2011, pp. 301–306.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition*, vol. 2, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005, pp. 886–893.

[17] A. Pobitzer, A. Lez, K. Matkovic, and H. Hauser, "A statistics-based dimension reduction of the space of path line attributes for interactive visual flow analysis," in *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis 2012)*, March 2012, pp. 113–120.

[18] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic huber-L^1 optical flow," in *British Machine Vision Conference (BMVC 09)*, 2009.

[19] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, oct. 2005, pp. 1395 –1402 Vol. 2.

[20] 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*, vol. 3, aug. 2004, pp. 32 – 36 Vol.3.

[21] (2005) Kth dataset split. [Online]. Available: http://www.nada.kth.se/cvap/actions/00sequences.txt