2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Optical Flow Co-occurrence Matrices: A Novel Spatiotemporal Feature Descriptor

Carlos Caetano, Jefersson A. dos Santos, William Robson Schwartz
Smart Surveillance Interest Group, Computer Science Department
Universidade Federal de Minas Gerais, Minas Gerais, Brazil
Email: {carlos.caetano,jefersson,william}@dcc.ufmg.br

***Abstract*—Suitable feature representation is essential for performing video analysis and understanding in applications within the smart surveillance domain. In this paper, we propose a novel spatiotemporal feature descriptor based on co-occurrence matrices computed from the optical flow magnitude and orientation. Our method, called Optical Flow Co-occurrence Matrices (OFCM), extracts a robust set of measures known as Haralick features to describe the flow patterns by measuring meaningful properties such as contrast, entropy and homogeneity of co-occurrence matrices to capture local space-time characteristics of the motion through the neighboring optical flow magnitude and orientation. We evaluate the proposed method on the action recognition problem by applying a visual recognition pipeline involving bag of local spatiotemporal features and SVM classification. The experimental results, carried on three well-known datasets (KTH, UCF Sports and HMDB51), demonstrate that OFCM outperforms the results achieved by several widely employed spatiotemporal feature descriptors such as HOF, HOG3D and MBH, indicating its suitability to be used as video representation.**

I. Introduction

Over the last decade, a significant portion of the progress in visual recognition tasks (e.g., object, action, and activity recognition) has been achieved with the design of discriminative local and global feature descriptors. In general, such representations are based on 2D or spatiotemporal feature descriptors employed for the image and video domains, respectively [1].

The feature extraction process is very important since it allows the image/video content to be represented in a more discriminative and compact space, when compared to the direct use of pixels. These representations must be rich enough to allow proper recognition. Typically, extracted features tend to be invariant to transformations such as rotation, scaling or changes on illumination. To that end, the most common is to extract 2D local feature descriptors of the image domain by using methods such as Scale Invariant Feature Transform (SIFT) [2], Speeded Up Robust Feature (SURF) [3], and Histogram of Oriented Gradients (HOG) [4].

The most employed local spatiotemporal feature descriptors are either generalizations of the aforementioned descriptors or based on motion analysis using optical flow [5]. Willems et al. [6] proposed an extension of the SURF feature know as eSURF, based on sums of Haar-wavelets in each cell. Laptev et al. [7] extracted a combination of local HOG and Histogram of Oriented Flow (HOF) descriptors from a set of interest points using a space-time extension of the Harris operator. An extension of HOG was presented by Kläser et al. [8], where 3D gradients are binned into regular polyhedrons using 3D integral images to allow rapid dense sampling of the cuboid over multiple scales and locations in both space and time. Another extension, the Gradient Boundary Histograms (GBH), was presented by Shi et al. [37] which is built on simple spatiotemporal gradients. The SIFT descriptor was also extended to its spatiotemporal version by Scovanner et al. [9].

There exists a large body of works on local spatiotemporal feature descriptors based on optical flow [10], [11], [12]. However, according to Oliva and Torralba [13], these descriptors have a performance well below of the desirable in extreme cases. Moreover, the flow information is not fully exploited due dimensionality reduction and histogram binning [14]. Consequently, these approaches do not necessarily capture interesting interactions from a surveillance point of view, discarding important information concerning spatial relations among the optical flow field [15].

Even though fewer, there are also works focusing on the design of global spatiotemporal feature descriptors. Mota et al. [16] presented a tensor motion descriptor using optical flow and HOG3D information. They consider a technique based on an aggregation tensor, combining two descriptors, in which one carries polynomial coefficients to approximate the optical flow and the other carries data from HOG. Recently, Shao et al. [17] presented a novel global descriptor for holistic human action recognition called spatiotemporal Laplacian pyramid coding (STLPC). This descriptor treats a video sequence as a whole and uses a 3D Gabor filter to make the features invariant to noise and distortion, followed by max pooling. As a drawback of global features is that it can be influenced by motions of multiple objects and variations in the background [18].

A recent growing trend is the employment of deep neural networks (DNNs) to feature learning [19], [20]. These architectures learn hierarchical layers of representations to perform pattern recognition and have demonstrated impressive state-of-the-art results on many pattern recognition tasks. However, a recent study by Nguyen et al. [21] revealed that state-of-the-art DNNs can be fooled, e.g., images that are completely unrecognizable by humans are predicted as recognizable objects by DNNs with over 99% of confidence, indicating that there is still room for feature design and engineering.

Aiming at capturing more information from the optical flow, this work proposes a novel spatiotemporal local feature de-

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

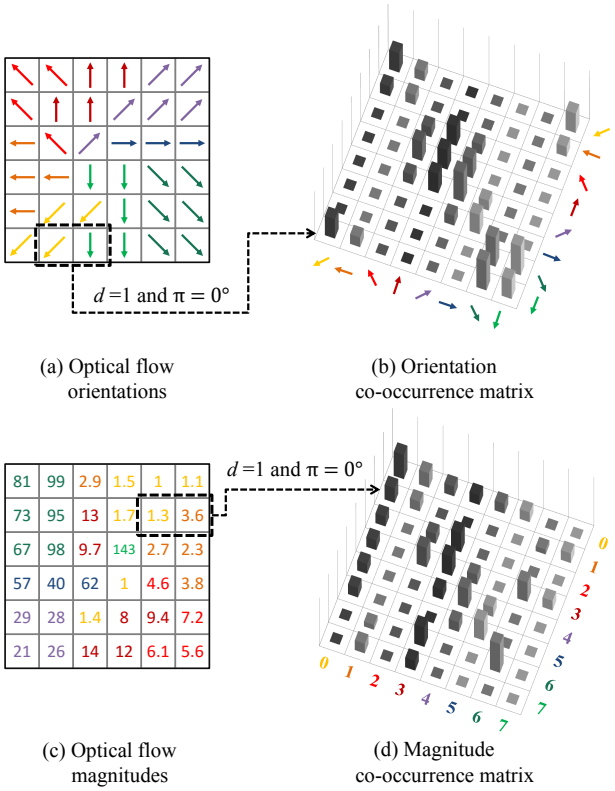


Fig. 1. Creation of the co-occurrence matrices from the orientation and magnitude information provided by the optical flow. For this example, the comparison between pairs of pixels is performed considering a 0° (horizontal) neighbor with distance $d = 1$ (for more details, see Section II). Maps in (a) and (c) represent a region of a frame considering orientations and magnitudes derived from the flow field, respectively. (b) and (d) represent the co-occurrence matrices computed from the optical flow orientation and magnitude, respectively (the latter considers a logarithm quantization).

scriptor called *Optical Flow Co-occurrence Matrices* (OFCM). The method is based on the extraction of Haralick features [22] from co-occurrence matrices computed using the optical flow information. In the literature, co-occurrence is often used to extract information related to global structures in various local region-based features, for instance Co-HOG [23], GLAC [24], STACOG [25] and CoF [26]. Moreover, co-occurrence is capable of extracting structural similarities leading to better performance [27]. Similar to our approach, the work of Maki et al. [26] proposed a pedestrian detection feature called Co-occurrence Flow (CoF) that encodes the unique motion of walking into the feature through pairwise comparisons of histograms of optical flow (HOF) for the entire body, i.e. across exhaustive combinations of cells.

Our hypothesis for designing the OFCM is based on the assumption that the motion information on a video sequence can be described by the spatial relationship contained on local neighborhoods of the flow field. More specifically, we assume that the motion information is adequately specified by a set of magnitude and orientation co-occurrence matrices computed for various angular relationships at a given offset between neighboring vector pairs on the optical flow. Therefore, matrices obtained by modifying the spatial relationship (different angles or distances between magnitudes and orientations of neighboring pixels from which the optical flow has been extracted) will provide different information. Figure 1 shows an example of co-occurrence matrices computed based on the orientations and magnitudes per pixel provided by the optical flow for an horizontal displacement of a single pixel between the frames. The final feature descriptor is obtained by the concatenation of the Haralick features extracted from each matrix.

Even though the method proposed by Maki et al. [26] is called Co-occurrence Flow, it does not compute a co-occurrence matrix on its description process. Moreover, the main difference between our work and [26] is that instead of using pairwise comparisons of HOF features, we consider the co-occurrence of magnitudes and orientations of the raw optical flow and, different from [26], we compute the Haralick features from the co-occurrence matrices while their final feature vector is composed by pairwise comparison using the $L1$ norm as a measure.

To demonstrate the effectiveness of the OFCM, we evaluate it in the action recognition task, a challenging problem that has attracted the attention of the research community for several years due to its practical and real-world applications [5], [28]. For instance, it can be employed on applications such as surveillance systems to detect and prevent abnormal or suspicious activities [29] and on health care systems to monitor patients performing activities of daily living [30]. Although we evaluate our method on the action recognition task, it is important to emphasize that, since the OFCM is a spatiotemporal feature descriptor, it can be also applied to other computer vision applications involving video description.

According to the experimental results, the proposed feature aggregated using a standard visual recognition pipeline (Bag-of-Words [31] followed by the SVM classifier) is able to recognize actions accurately on three well-know datasets: KTH, UCF Sports and HMDB51. The employment of the OFCM outperforms the results achieved by several widely employed spatiotemporal descriptors available in the literature.

II. Proposed Approach

Several spatiotemporal feature descriptors, such as [11] and [12], are based on gradient or optical flow to encode the information extracted from the video. While those works have demonstrated encouraging recognition accuracy, a rich source of information contained in these descriptors, such as spatial relations contained on the optical flow field, have not been fully explored.

To explore the local relations contained on the optical flow field, we propose a novel spatiotemporal feature descriptor, called Optical Flow Co-occurrence Matrices (OFCM), based on the co-occurrence matrices computed over the optical flow field. This co-occurrence matrices will express the distribution of the magnitude and orientation components at a given offset over the optical flow, as illustrated in Figure 1.

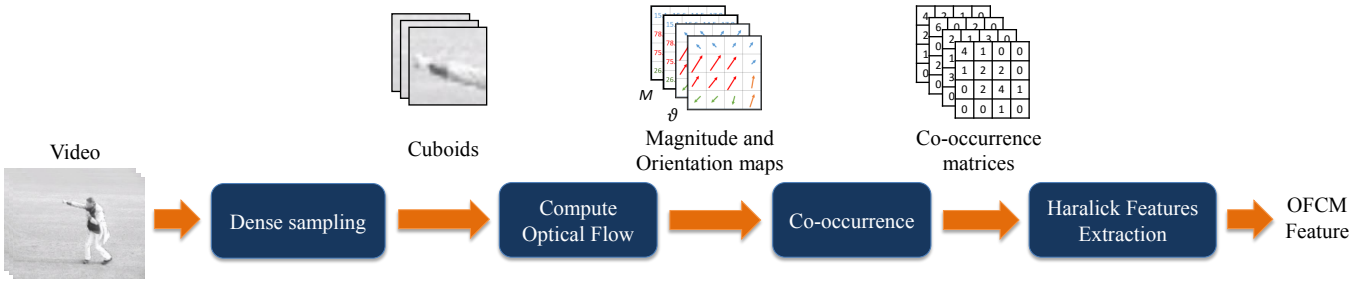


Fig. 2. Diagram illustrating the pipeline extraction of the proposed spatiotemporal feature descriptor.

Our hypothesis to design the OFCM is based on the assumption that the motion information on a video sequence can be captured by the overall relationship of the vectors in the optical flow. In addition, we believe that it can be specified by a set of magnitude and orientation dependence matrices computed for various angular relationships and distances between neighboring vector pairs on the video optical flow. Once the co-occurrence matrices have been computed, we use a set of measures known as Haralick textural features [22] to describe the flow patterns.

The classical textural feature gray level co-occurrence matrix (GLCM), proposed by Haralick et al. [22], estimates the joint distribution of pixel intensity given a distance and an orientation. The co-occurrence matrix Σ is defined over an $n \times m$ image I, at a specified offset (Δ_x, Δ_y), as

$$\Sigma_{\Delta_x,\Delta_y}(i,j) = \sum_{r=1}^{n}\sum_{q=1}^{m} \begin{cases} 1, & \text{if } I(r,q) = i \text{ and} \\ & I(r+\Delta_x, q+\Delta_y) = j, \\ 0, & \text{otherwise} \end{cases}$$

where i and j are the image intensity values of pixels separated by a distance d, r and q are the spatial positions in the image I and the offset (Δ_x, Δ_y) depends on the angle π used. Usually, π is expressed as angles 0° $(0,d)$, 45° $(-d,d)$, 90° $(-d,0)$ and 135° $(-d,-d)$. Fig. 3 illustrates possible offset configurations. Note that in the proposed method, we do not compute the matrices using the image intensity values, but using maps θ^Q e M^Q, which will be discussed in the next paragraphs.

The process of computing the OFCM is illustrated in Figure 2 and will be explained in details as follows. First, a dense sampling step is applied to the video dividing it into $n_i \times n_j \times n_t$ regions. These regions are referred to as cuboids and are described by their width (n_i), height (n_j), and length (n_t). With the cuboids at hand, we build an orientation-magnitude representation derived from the optical flow. Since extracting optical flow for every pixel is computationally expensive [15], we create a binary mask using absolute difference image between the frame I_t and the frame I_{t+k} and given a threshold h, if the resulting difference is less than h, the pixel is discarded; otherwise, this pixel p is set to its corresponding local cuboid C_i. We compute the optical flow using the Lucas-Kanade Pyramid [32].

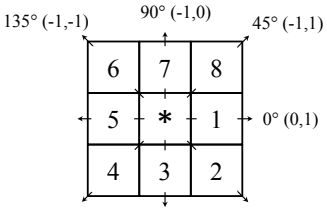


Fig. 3. Offset configurations with $d = 1$. Cells 1 and 5 are the 0° (horizontal) nearest neighbors to cell *; cells 2 and 6 are the 135° nearest neighbors; cells 3 and 7 are the 90° nearest neighbors; and cells 4 and 8 are the 45° nearest neighbors to *. Note that this information is purely spatial, and has nothing to do with pixel intensity values [22].

As mentioned earlier, OFCM uses optical flow information (orientation and magnitude) to build co-occurrence matrices for each cuboid. To this end, we build two maps for each flow field computed on the cuboid: one based on the orientations of the flow field and another one based on the magnitudes of the flow field. In this way, for each cuboid C_i composed by n_t frames, we compute $n_t - 1$ magnitude maps M and $n_t - 1$ orientation maps θ using Equations 1 and 2 defined as

$$M_{i,j} = \sqrt{u_{i,j}^2 + v_{i,j}^2} \tag{1}$$

and

$$\theta_{i,j} = tan^{-1}\left(\frac{v_{i,j}}{u_{i,j}}\right), \tag{2}$$

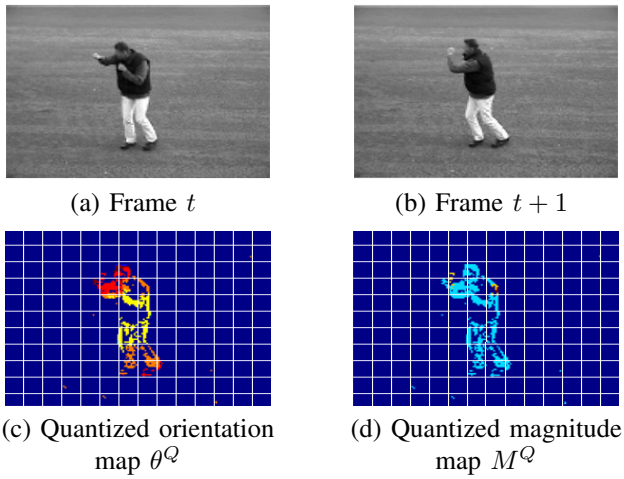
where u and v represent the horizontal and vertical components of each the flow vector contained in the flow field.

Since the values obtained in M and θ maps are composed by real numbers, a quantization step is applied to compute the co-occurrence matrices of M and θ. The magnitude quantization used is based on *dLog* distance proposed in [33]. Here, our main goal is to reduce the impact of noisy high magnitude values. The quantization functions are defined as

$$M_{i,j}^Q = \begin{cases} \sigma, & \text{if } m_{i,j} > \sigma \\ 0, & \text{if } m_{i,j} < 0 \\ m_{i,j}, & \text{otherwise} \end{cases} \tag{3}$$

$$m_{i,j} = \lfloor \log_2 M_{i,j} \rfloor \tag{4}$$

$$\theta_{i,j}^Q = \lfloor \theta_{i,j} / (\Theta/\omega) \rfloor, \tag{5}$$



(a) Frame t (b) Frame $t+1$

(c) Quantized orientation map θ^Q (d) Quantized magnitude map M^Q

Fig. 4. Quantized θ^Q and M^Q maps obtained from the flow filed between frames t and $t+1$. Best viewed in color.

where Θ is the maximum orientation value, σ and ω are the numbers of desired magnitude and orientation bins, respectively. Figure 4 illustrates the quantized θ^Q and M^Q maps between two frames. Figure 4 (c) illustrates a concatenation of all θ^Q maps extracted from a dense grid of cuboids between a frame t and frame $t+1$. The same is illustrated for the M^Q maps in Figure 4 (d).

For each cuboid, we compute α co-occurrence matrices, where α is the number of angles used ($\pi = 0°$, $45°$, $90°$ and $135°$), from the quantized M^Q and θ^Q maps according to Equation II. After that, we extract f Haralick textural features [22] for each co-occurrence matrix, generating a feature vector with f dimensions per matrix, where f is the number of extracted Haralick features. Finally, all feature vectors are concatenated, providing a final representation with length of $2 \times (\alpha \times f) \times n_t - 1$.

The source-code with the implementation of the OFCM spatiotemporal feature descriptor is available at SSIGLib[1].

III. Experimental Results

This section describes the experimental results obtained with the OFCM for the action recognition problem and compares it to other feature descriptors in the literature. Besides the well-known spatial HOG [4] descriptor, five widely employed spatiotemporal feature descriptors used by state-of-the-art approaches were chosen to be compared with our proposed feature: HOF [7], HOG/HOF [7], HOG3D [8], MBH [10] and GBH [37]. We also compare the proposed OFCM with the Dense Trajectories (DT) method [12], which employs the combination of three descriptors (HOG [4], HOF [7], and MBH [10]) in conjunction with Bag-of-Words and Support Vector Machines (SVM) as classifier.

To isolate only the contribution brought by the feature descriptors to the action recognition problem, all descriptors were tested on the same datasets with the same split of training and testing data and using the same classification method. We evaluate the features in a Bag-of-Words-based action classification task and employ the evaluation protocols and metrics proposed by the creators of the datasets.

A. Datasets

To evaluate the proposed feature descriptor, we consider three well-known datasets for the action recognition problem, KTH, UCF Sports and HMDB51.

The *KTH actions dataset* [18] consists of six human action classes. Each action is performed several times by 25 subjects on four different scenarios. In total, the data consist of 600 videos with frame rate of 25 fps and spatial resolution of 160×120 pixels. To obtain fair comparison, we follow the experimental setup used by Wang et al. [34] and divide the samples into training/validation set ($8+8$ people) and test set (9 people). The performance is evaluated as suggested in [18], i.e., by reporting the average accuracy over all classes.

The *UCF sports* is a realistic dataset [35] composed by ten different types of human actions collected from various sports. It consists of 150 video samples with a frame rate of 10 fps and spatial resolution of 720×480 pixels. It is a very challenging dataset presenting a large intraclass variability and variations in camera motion. We use a leave-one-out setup and average accuracy over all classes as performance metric, as suggested by the authors [35].

The *HMDB51* [36] is realistic and challenging action dataset composed of video clips from movies, the Prelinger archive, Internet, Youtube and Google videos, and comprised of 51 action categories. It consists of 6766 action samples with a resolution of 240 pixels in height with preserved aspect ratio. We follow the original protocol using three train-test splits. The performance is evaluated by computing the mean accuracy over the three splits as suggested by the authors [36].

B. Recognition Pipeline

Aiming a fair comparison, we apply the same evaluation pipeline for every feature descriptor as in [12], [34]: a classical visual recognition pipeline involving training and test steps.

In the training phase, we first densely extract spatiotemporal feature descriptors. Dense sampling extracts video blocks at regular positions in space and time. There are 3 dimensions to sample from: $n_i \times n_j \times n_t$. In our experiments, the default size of a block is 18×18 pixels with 10 frames and a 50% of overlapping on both spatial and temporal sampling. Then, following the visual recognition strategy, the local features are encoded into a mid-level representation to be used for the classification task. However, a visual codebook must be created before the encoding. Therefore, we randomly sample V training features. This is very fast and according to [8] the results are very close to those obtained using vocabularies built with k-means. We set the number of visual words V to 4000 which, according to [34], has shown to achieve good results for a wide range of datasets. Afterwards, for each video sequence, we compute a Bag-of-Words feature vector. Spatiotemporal features are first quantized into visual words and a video is

[1] https://www.github.com/ssig/ssiglib

Cuboid spatial size variation	KTH Acc. (%)	UCF Sports Acc. (%)
OFCM ($n_i = 18, n_j = 18, n_t = 10$)	95.83	**92.80**
OFCM ($n_i = 24, n_j = 24, n_t = 10$)	95.99	87.73
OFCM ($n_i = 36, n_j = 36, n_t = 10$)	**96.30**	90.00
OFCM ($n_i = 48, n_j = 48, n_t = 10$)	95.83	89.73
OFCM ($n_i = 72, n_j = 72, n_t = 10$)	95.52	90.27

TABLE I
ACTION RECOGNITION ACCURACY (%) RESULTS OF OFCM. CUBOID SPATIAL SIZE VARIATION ON KTH [18] AND UCF SPORTS [35] ACTION DATASETS.

then represented as the frequency histogram over the visual words. Euclidean distance is applied as the distance metric between the features and the closest vocabulary word. Finally, an one-against-all classification is performed by a non-linear Support Vector Machines (SVM) with a RBF-kernel.

In test phase, a test video sequence is classified by applying the trained classifier obtained during the training phase. Therefore, for a test video sequence, spatiotemporal feature descriptors are extracted with dense sampling. Then, the Bag-of-Words feature vector is generated using the visual codebook previously created. Finally, the generated feature vector is given as input to the trained classifier to predict the class label of the test video sequence.

It is important to emphasize that in the experiments, we only change the feature descriptor used in the pipeline since our main goal is to compare the real contribution of our proposed feature descriptor, i.e., for each experiment, the pipeline is the same, only the feature descriptor is switched.

C. Parameter Setting

This section presents experiments regarding parameter setting for the OFCM focusing on the optimization of the number of bins for optical flow magnitude and orientation, the offset distance d, used to create the co-occurrence matrix, and the spatial size of the cuboid. We extract $f = 12$ Haralick textural features for all the following experiments.

We optimized the number of magnitude and orientation bins parameters (σ and ω), on KTH and UCF Sports datasets, respectively. Here, we empirically set the offset distance $d = 1$. On the KTH dataset, the best result (95.83%) was achieved with both $\omega = 8$ and $\sigma = 4$. On the other hand, the best result (92.00%) for UCF Sports dataset was achieved with $\omega = 4$ and $\sigma = 8$. On the UCF Sports dataset, the best result was obtained with a higher number of magnitude bins parameters (σ) than on KTH dataset. We believe this is partly because UCF Sports videos are composed of a lower frame rate (10 fps) presenting higher magnitude values.

We also optimized the offset distance d parameter. For this purpose, we fixed σ and ω with the best parameters obtained on the previous experiments. In our test, we varied it to a maximum of $d = 5$. On the KTH dataset, the best result maintained with $d = 1$ (95.83%). However, on UCF Sports dataset, the best result was achieved with $d = 5$ (92.80%).

Table I reports the performance of different spatial size of the cuboid. As in [34], we fixed the temporal length n_t to 10 frames and an overlapping rate of 50%. On KTH dataset, the best result (96.30%) was achieved at a spatial size of 36×36 pixels. On the other side, the best result for UCF Sports dataset was achieved with a spatial size of 18×18 pixels (92.80%).

D. Results and Comparisons

Now, we compare our approach with several classic local spatiotemporal features of the literature. According to Table II, a considerable improvement was obtained with OFCM, reaching 96.30% of accuracy on KTH dataset and 92.80% on UCF Sports dataset. There is an improvement of 2.10 percentage points (p.p.) on the KTH dataset and 3.80 p.p. on the UCF Sports dataset achieved by the OFCM when compared to Wang et al. DT method [12]. Furthermore, it is worth noting that their approach uses a combination of three different feature descriptors (HOG, HOF and MBH) while we only used our proposed OFCM feature. Therefore, such results can be considered remarkably good and confirm the advantages introduced by our spatiotemporal feature descriptor.

For the experiments on HMDB51 dataset, we used the parameters learned using the KTH dataset, as they turned out to be universal enough to obtain accurate results [8]. For this dataset the OFCM also achieves the best results, reaching 56.91% of accuracy, an improvement of 5.41 p.p. when compared to the MBH feature descriptor [10]. This results demonstrate the generalization ability of the OFCM once its parameters were estimated using data from the KTH.

The improvement of OFCM over the other descriptors is especially striking. We believe the reason for this is that our spatiotemporal feature is capturing important temporal motion information since a pair (co-occurrence) of magnitudes or orientations have more "vocabulary" than a single histogram bin of gradient, magnitude or orientation. In this way, the OFCM can express motion in more details than the other histogram based features, which use single gradient orientation or single flow orientations discarding important information concerning spatial relations among the flow field.

IV. CONCLUSIONS AND FUTURE WORK

In this work, we proposed a novel spatiotemporal feature descriptor called Optical Flow Co-occurrence Matrices (OFCM). The method is based on the extraction of Haralick features from the co-occurrence matrices derived of the orientation and magnitude from the vector field obtained by optical flow. The co-occurrence matrices are capable of extracting structural similarities of the motion information by the distribution of the magnitude and orientation components at a given offset over the optical flow and the Haralick features are responsible for measuring statistical properties such as homogeneity, entropy, linear dependencies (linear structure), number and nature of boundaries present, and the complexity of the flow field. We demonstrated that OFCM outperforms several classic spatiotemporal features of the literature on three well-known

	Approach	KTH Acc. (%)	UCF Sports Acc. (%)	HMDB51 Acc. (%)
Published results	HOG [4]	79.00	77.40	28.40
	HOF [7]	88.00	84.00	35.50
	HOG/HOF [7]	86.10	81.60	43.60
	HOG3D [8]	85.30	85.60	36.20
	MBH [10]	89.04	90.53	51.50
	GBH [37]	92.70	-	38.80
	DT [12]	94.20	88.20	46.60
Our results	OFCM	**96.30**	**92.80**	**56.91**

TABLE II
ACTION RECOGNITION ACCURACY (%) RESULTS OF OFCM AND CLASSIC SPATIOTEMPORAL FEATURES OF THE LITERATURE ON KTH [18] AND UCF SPORTS [35] ACTION DATASETS. RESULTS FOR HOG, HOF, HOG/HOF AND HOG3D WERE OBTAINED FROM [34].

action recognition datasets, KTH, UCF Sports and HMDB51 (by up to 2.10 p.p., 3.80 p.p., and 5.41 p.p., respectively).

Directions for future works include the evaluation of the proposed spatiotemporal feature on different action recognition pipelines and investigate the impact of each Haralick feature on the flow field. Moreover, we intend to evaluate the proposed feature descriptor in other action datasets as well as evaluate its behavior on other video-related problems.

ACKNOWLEDGMENTS

The authors would like to thank the Brazilian National Research Council – CNPq (Grant #477457/2013-4), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00025-15) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

REFERENCES

[1] S. Krig, "Interest point detector and feature descriptor survey," in *Computer Vision Metrics*, 2014.
[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.
[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *CVIU*, vol. 110, pp. 346–359, 2008.
[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
[5] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, pp. 976–990, 2010.
[6] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV*, 2008.
[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.
[8] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008.
[9] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th International Conference on Multimedia*, 2007.
[10] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, 2006.
[11] I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition," in *Spatial Coherence for Visual Motion Analysis*. Springer Berlin Heidelberg, 2006.
[12] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.
[13] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, pp. 520 – 527, 2007.
[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *PETS*, 2005.
[15] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Textures of optical flow for real-time anomaly detection in crowds." in *AVSS*, 2011.
[16] V. Mota, E. Perez, L. Maciel, M. Vieira, and P. Gosselin, "A tensor motion descriptor based on histograms of gradients and optical flow," *Pattern Recognition Letters*, vol. 39, pp. 85 – 91, 2014.
[17] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *Cybernetics, IEEE Transactions on*, vol. 44, pp. 817–827, 2014.
[18] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *ICPR*, 2004.
[19] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *TPAMI*, vol. 35, pp. 1915–1929, 2013.
[20] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised multi-stage feature learning," in *CVPR*, 2013.
[21] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *CVPR*, 2015.
[22] R. M. Haralick, K. S. Shanmugam, and I. Dinstein, "Textural features for image classification." *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, 1973.
[23] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence histograms of oriented gradients for pedestrian detection," in *PSIVT*, 2008.
[24] T. Kobayashi and N. Otsu, "Image feature extraction using gradient local auto-correlations," in *ECCV*, 2008.
[25] ——, "Motion recognition using local auto-correlation of space-time gradients," *Pattern Recogn. Lett.*, vol. 33, pp. 1188–1195, 2012.
[26] A. Maki, A. Seki, T. Watanabe, and R. Cipolla, "Co-occurrence flow for pedestrian detection," in *ICIP*, 2011.
[27] T. Mita, T. Kaneko, B. Stenger, and O. Hori, "Discriminative feature co-occurrence selection for object detection," *TPAMI*, vol. 30, pp. 1257–1269, 2008.
[28] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *CVIU*, vol. 115, pp. 224–241, 2011.
[29] A. Wiliem, V. Madasu, W. Boles, and P. Yarlagadda, "A suspicious behaviour detection using a context space model for smart surveillance systems," *CVIU*, vol. 116, pp. 194–209, 2012.
[30] C. Zhang and Y. Tian, "Rgb-d camera-based daily living activity recognition," *IJCVIP*, vol. 2, 2012.
[31] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.
[32] J. yves Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.
[33] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *CIKM*, 2002.
[34] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.
[35] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.
[36] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *ICCV*, 2011.
[37] F. Shi, R. Laganiere, and E. Petriu, "Gradient boundary histograms for action recognition," in *WACV*, 2015.