Pyramidal Zernike Over Time: A Spatiotemporal Feature Descriptor Based on Zernike Moments

Igor L. O. Bastos[1(✉)], Larissa Rocha Soares[2], and William Robson Schwartz[1]

[1] Smart Surveillance Interest Group, Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
igorcrexito@gmail.com, william@dcc.ufmg.br

[2] Reuse in Software Engineering, Universidade Federal da Bahia, Salvador, Brazil
larissars@dcc.ufba.br

Abstract. This paper aims at presenting an approach to recognize human activities in videos through the application of Zernike invariant moments. Instead of computing the regular Zernike moments, our technique, named Pyramidal Zernike Over Time (PZOT), creates a pyramidal structure and uses the Zernike response at different levels to associate subsequent frames, adding temporal information. At the end, the feature response is associated to Gabor filters to generate video descriptions. To evaluate the present approach, experiments were performed on the UCFSports dataset using a standard protocol, achieving an accuracy of 86.05%, comparable to results achieved by other widely employed spatiotemporal feature descriptors available in the literature.

Keywords: Activity recognition · Feature extraction
Zernike moments

1 Introduction

Over the previous years, a growing interest in video surveillance applications such as elder care, home nursing and video monitoring, has been noticed, encouraging researches related to the field of activity recognition [3]. These activities are connected with the time domain and important information is denoted by changes in appearance and motion of elements over time [8]. In a simplistic way, they are sequences of primitive sub-actions [3], in which spatial and temporal features are associated to describe the whole event.

To describe activities in videos, features are extracted allowing a representation in a discriminative and compact way [2]. Due to the richness of spatial information on every frame, 2D local feature descriptors are commonly applied in this field. However, these descriptors do not consider the time contribution for this task, making room for the use of spatiotemporal descriptors. In this context, this work exploits Zernike Moments [9] and proposes a feature descriptor, called

M. Mendoza and S. Velastín (Eds.): CIARP 2017, LNCS 10657, pp. 77–85, 2018.
https://doi.org/10.1007/978-3-319-75193-1_10

Pyramidal Zernike Over Time (PZOT), that takes into account the temporal information in the description video sequences containing human activities.

This paper proposes a novel spatiotemporal method to recognize activities in videos based on a descriptor derived from the conventional Zernike invariant moments. To evaluate the method, experiments are performed on the UCFSports dataset, employing a standard visual recognition pipeline (Bag-of-Words [17] followed by the SVM classifier), reaching an accuracy of 86.05%, which is comparable to results achieved by other widely employed spatiotemporal feature descriptors, demonstrating the viability of the proposed approach.

2 Related Work

Due to the recent interest and large demand for activity recognition approaches, many studies have been developed for this task. The main idea of using Zernike invariant moments on the present approach regards the importance of spatial information, making this technique suitable to recognize activities.

The application of Zernike moments derives from their ability to provide information related to shapes in images. This fact was exploited by Tsolakidis et al. [18] and Newton and Hse [13], where the technique was used for leaf and digit recognition, respectively. Besides, the use of Zernike moments is desirable due to their amplitude invariance to rotation and translation, the possibility of changing their order and repetition parameters, and their ability to extract finer or coarser information regarding shapes with no overlap or redundancy [9].

The present approach tackles the extension of conventional Zernike Moments with the aim of adding temporal description. In this sense, the approaches proposed by Lassoued et al. [12] and Constantini et al. [4] deserve attention, since both designed Zernike based methods that deal with time information. Lassoued et al. [12] presented a temporal adaptation of Zernike moments to recognize human activities consisting on the extraction of 3D silhouettes over time on which a variant of Zernike moments is applied. Although it also adapts the Zernike moments to the temporal context, the work proposed in [12] differs from the present approach since it proposes a volumetric modeling of moments, changing their formulation to contemplate time. Similarly, the method proposed by Constantini et al. [4] is based on a volumetric modeling of Zernike moments computed over subsequent frames and associated to a spatial pyramid for matching.

Although presenting similarities with the aforementioned Zernike-based methods, the current approach presents a different idea since the temporal information is gathered from the combination of 2D Zernike responses. Moreover, the pyramid applied on the proposed approach is different from the one proposed in [4], since in [4] it consists of an arrangement of an image in different spatial scales, which are considered for matching descriptor responses. In turn, here, we create a pyramid by splitting the image in an increasing number of grids, used to extract the finer shape information through the Zernike responses.

PZOT and the methods proposed in [4] and [12] modify a 2D spatial descriptor to incorporate temporal information. This strategy is followed by well-known

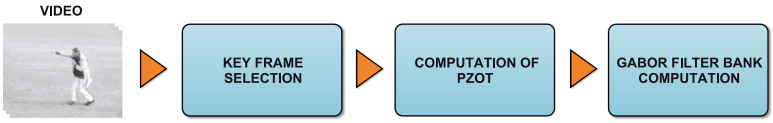


Fig. 1. Steps of the proposed approach.

approaches, producing spatiotemporal descriptors from spatial descriptors, such as HOG3D [10] or the temporal strands of SURF [20] and SIFT [15].

Nowadays, one can notice the increasing number of deep neural models used for activity recognition, as the ones proposed by Simonyan and Zisserman [16] and Feichtenhofer et al. [7]. However, this paper focuses on the proposition of a handcrafted feature descriptor. These descriptors are valuable since they can be associated to less complex classification models, which present a lower computational cost for training and require a smaller amount of training data. Hence, the validation performed in this study is based on tests and comparisons between the present method and state-of-art spatiotemporal handcrafted descriptors.

3 Proposed Approach

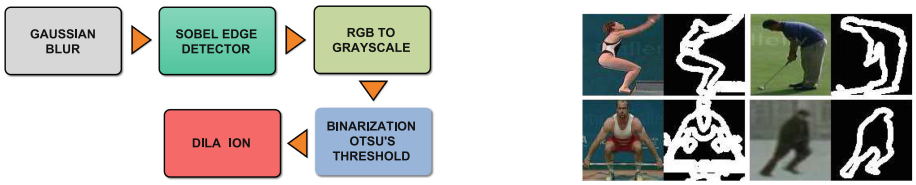
The present research proposes a novel spatiotemporal feature descriptor based on the usage of Zernike invariant moments to be applied to the activity recognition task. Thus, to operate correctly, some steps need to be taken, ranging from an initial selection of key frames in the video to the feature extraction with Zernike and Gabor descriptors. This section describes the steps of our approach, illustrated in Fig. 1. It is important to mention that the descriptor proposed in this work is intended to be applied over bounding boxes of the subjects performing the activity in the scene, usually provided by the video datasets.

3.1 Key Frame Selection

The first step of the approach performs a key frame selection to find relevant frames and avoid redundancy and unnecessary processing. For every video, the difference between subsequent frames is computed and those that satisfy a threshold (empirically determined) are selected as relevant. This strategy is based on the idea that similar frames do not contribute for the activity recognition. The feature extraction performed on PZOT occurs only on the key frames.

3.2 Computation of Pyramidal Zernike Over Time

To obtain information related to shapes on frames, a sequence of image processing techniques is employed before computing the Zernike moments. This step intends to produce binary masks in which the subject contours (shape) are highlighted. In addition, internal contours tend to be discarded, resulting in images



(a) (b)

Fig. 2. Preprocessing steps prior to Zernike moments computation. (a) Sequence of image processing techniques; (b) resulting binary masks.

containing silhouettes of the subjects. With that, the Zernike response for each pair of parameters regards these silhouettes and, consequently, produces a more accurate outcome, which is more related to subject postures and is less influenced by background. This sequence of image processing techniques must be applied on every bounding box region of every frame of the videos. Figure 2(a) illustrates the processing pipeline and Fig. 2(b) exemplifies some obtained outputs. It is worth mentioning that both Sobel and Gaussian blur operators are applied over RGB images considering each channel independently. Furthermore, the dilation operation is performed over binary images using a 4×4 disk structuring element.

Instead of computing regular Zernike moments, a temporal approach, named the Pyramidal Zernike Over Time (PZOT) is assembled. This approach consists of placing every selected frame in different pyramid levels, in which every level is represented by the original image split in an increasing number of subimages. The number of subimages ranges from 1 (level one at the top of pyramid) to 4^{n-1} (at the base of pyramid), where n represents the number of levels.

At every level of the pyramid, nine Zernike moments are computed (amplitude and phase) for each subimage. These moments differ one from another according to the order and repetition parameters, making them able to represent finer and coarser details of the shapes. The Zernike moments are obtained from

$$Z_{nm} = \frac{(n+1)}{\pi} \sum_{x} \sum_{y} f(x,y) V_{nm}(x,y), \tag{1}$$

where $V_{nm}(x,y)$ represents the complex Zernike polynomial of order n and repetition m projected in polar coordinates and $f(x,y)$ represents a pixel of the input image. The phase and amplitude are obtained from each value of Z_{nm}.

The goal of splitting in an increasing number of subimages is to capture local shape information, obtaining responses to shapes of the entire image and local regions, improving the descriptiveness of the technique. For a three-level pyramid, for instance, $9 \times (1) + 9 \times (4) + 9 \times (16)$ moments are computed, resulting in 189 amplitude and 189 phase values.

Despite the increase of descriptiveness, the pyramid structure does not incorporate temporal information. To this end, it is necessary to create a relation between the frames of the video, which is made by the difference of the amplitude and phase values of consecutive frames, generating coefficients that describe

distortions in the shapes (amplitude difference) and their positioning (phase difference). Therefore, the PZOT descriptor is composed by the computation of Zernike Pyramid for the first frame (representing the base shape) and the difference of every pyramid level for each subsequent pair of frames, resulting in a descriptor with 189 (base frame) $+ (k - 1) \times 189$ values for both amplitude and phase, where k represents the number of video frames. Section 4 describes a variation of this strategy that produced better results, in which Zernike responses are computed for each frame and added to this feature descriptor.

3.3 Gabor Filter Bank Computation

PZOT describes a video sequence through the base shape and its variations on subsequent frames. This strategy efficiently computes these variations, but does not capture much information to describe each frame, not exploring an important source of information for the task of activity recognition. Thus, Gabor filters have been included on the approach to describe the contours at different frequencies and orientations. These filters were intended to be applied over bounding boxes provided by video datasets (not applied on binary masks). However, instead of computing Gabor responses for every frame, we only compute for frames where the amplitude difference of Zernike moments satisfy a threshold value, indicating a strong variation on the silhouette (appearance) of the action subject.

To select the frames from which Gabor filters will be extracted, the Zernike amplitudes are used to compose a curve at the first level of pyramid for each frame. The difference of areas under these curves for a frame and the base shape frame are compared to a threshold (20% of the area under the base shape curve). Every time this value is satisfied, the frame is considered relevant (different from its predecessor) and the Gabor responses are computed. It is important to mention that the difference is computed, initially, according to the first frame (first base shape). Afterwards, this base shape is updated every time the threshold is satisfied and every subsequent comparison is performed considering this update. Figure 3(a) shows the responses for two subsequent frames for a diving activity video (red and blue curves) of the UCFSports dataset.

An 8-degree polynomial (derived from nine amplitude responses of Zernike moments) is applied to adjust the curves and after that, the difference is computed (showed in yellow in Fig. 3). The degree of the polynomial enforces a non-linear relation among amplitude values, which implicitly corresponds to responses for different frequencies. Besides that, fitting a curve emphasizes the difference between responses for subsequent frames. The degree of this polynomial was adjusted experimentally. For any degree smaller than 8, the curves were not well adjusted to the 9 Zernike amplitudes, resulting in low accuracies.

The Gabor filter bank employed consists of filters with four different orientations and four different frequencies. Those filters are weighted according to the difference obtained in the curve. Figure 3(b) shows a curve difference obtained for two non-consecutive frames of a diving activity video of the UCFSports dataset. It is possible to notice that the difference between curves (showed in light and dark yellow) is more relevant at the high order moments, which indicate

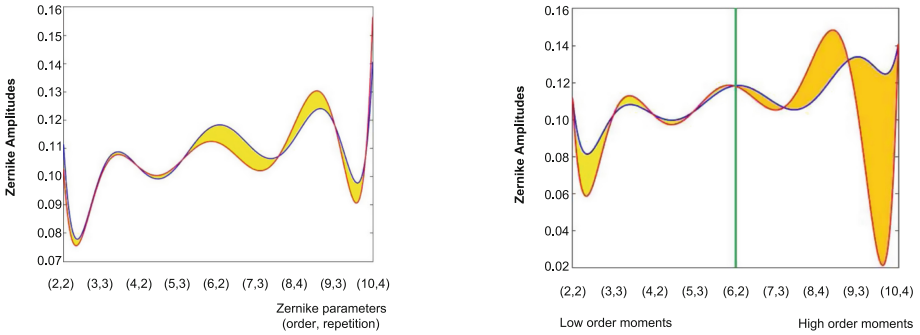


Fig. 3. Difference between Zernike curves (amplitude as a function of order/repetition). Both curves were fit by a 8-degree polynom. (a) Difference for two subsequent frames; (b) difference for two non-subsequent frames. (Color figure online)

a stronger variation on finer shapes. Such Zernike behavior may be associated to high frequency responses, as low order moments may be associated to low frequency responses. Thus, the curve difference is split into two regions and the Gabor filters with the two lowest frequencies are weighted by the low order area, while the Gabor filters with the two highest frequencies are weighted by the high order area, as shown on Fig. 3(b). The weighting coefficients are computed by the area of each region normalized by the total difference area.

4 Experimental Results

The UCFSports dataset [14] was used to evaluate the approach. This dataset is composed by 150 sequences representing 10 different sport activities. It also provides, along with videos, bounding boxes detaching the subjects involved in each activity. Furthermore, to homogenize the tests, a classification pipeline described in [2] was employed. It uses SVM classifier with a RBF kernel and a bag of 4000 visual words. The executed tests intend to evaluate the feasibility of our descriptor, even if the obtained results depend on steps that precede its use.

The first step of the evaluation of our method regards the selection of Zernike parameters, which were chosen by a preliminary study conducted on a Brazilian Sign Language recognition dataset [1]. From this study, nine pairs of parameters (order and repetition) were set, varying from lower to higher order moments: (10, 4), (9, 3), (8, 4), (7, 3), (6, 2), (5, 3), (4, 2), (3, 3), and (2, 2).

Once the Zernike parameters have been defined, we conducted a test with the application of PZOT (without Gabor filters) on the video frames, computing the base shape for the first frame and amplitude and phase distortions over time. We achieved an accuracy of 63.64%. Then, a second test was conducted. Instead of computing only the distortions regarding the base shape, the Zernike amplitude/phase responses (describing shapes of every frame) were added, reaching an accuracy of 70.54%. Finally, the complete approach was tested (adding Gabor), as described in Sect. 3.3. This test achieved an accuracy of 86.05%.

Table 1. Accuracies of different approaches applied to the UCFSports dataset.

	Approach	Acc (%)
Published results	HOG [5]	77.00
	HOF [11]	84.00
	HOG/HOF [11]	81.60
	HOG3D [10]	85.60
	MBH [6]	90.53
	DT [19]	88.20
	OFCM [2]	92.80
Our results	PZOT with base shape and only distortions	63.64
	PZOT with shapes on every frame	70.54
	Gabor with no PZOT	79.47
	PZOT + Gabor (proposed approach)	**86.05**

Fig. 4. Binary masks with noisy shapes detached.

According to Table 1, PZOT provides results that are comparable to previous researches applied to the UCFSports dataset. It is important to highlight that the features extracted with PZOT contribute with the ones extracted by Gabor filters in a complementary way, what is revealed by the increase of accuracy obtained with the combination of techniques. Lastly, three pyramid levels were applied, as they produced better results than two or four levels.

The *Skateboarding* and *Running* classes of UCFSports were the ones for which the method obtained the lowest accuracies: 75.09% and 70.91%, respectively. For these classes, the image processing pipeline generated noisy images, preventing the Zernike moments from being properly extracted, as shown in Fig. 4. One can notice the amount of clutter and how human shapes are not well-segmented.

5 Conclusions

This paper presented a temporal feature descriptor based on the Zernike invariant moments. Tests were conducted to evaluate this descriptor on the UCFSports dataset, reaching an accuracy of 86.05%. Despite achieving a smaller accuracy than some Spatiotemporal descriptors, the present approach shows to be valuable since it achieved comparable results using a simple and inexpensive method. For some activities of the UCFSports dataset, the number of extracted features is approximately 1/3 of the ones extracted by the OFCM [2], the state-of-art

feature descriptor for this dataset. At last, studies should be directed to points such as the simplicity of the image processing pipeline, application of few Zernike pairs of parameters and evaluation of a different Gabor filter bank, representing future directions for the present approach.

Acknowledgments. The authors would like to thank the Brazilian National Research Council – CNPq, the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

References

1. Bastos, I.L.O., Angelo, M.F., Loula, A.: Recognition of static gestures applied to Brazilian sign language (libras). In: SIBGRAPI, pp. 305–312 (2015)
2. Caetano, C., Santos, J.A., Schwartz, W.R.: Optical flow co-occurrence matrices: a novel spatiotemporal feature descriptor. In: Proceedings of the 23rd ICPR (2016)
3. Chaquet, J.M., Carmona, E.J., Antonio, F.: A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. **117**, 633–659 (2013)
4. Costantini, L., Seidenari, L., Serra, G., Capodiferro, L., Del Bimbo, A.: Space-time Zernike moments and pyramid kernel descriptors for action classification. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part II. LNCS, vol. 6979, pp. 199–208. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24088-1_21
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE CVPR 2005, vol. 1, pp. 886–893 (2005)
6. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006). https://doi.org/10.1007/11744047_33
7. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE CVPR 2016 (2016)
8. Ke, S., Thuc, H., Lee, Y., Hwang, J., Yoo, J., Choi, K.: A review on video-based human activity recognition. Computers **2**, 88–131 (2013)
9. Khotanzad, A., Hong, Y.H.: Invariant image recognition by Zernike moments. IEEE Trans. Pattern Anal. Mach. Intell. **12**(5), 489–497 (1990)
10. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of 19th BMVC, pp. 275:1–10 (2008)
11. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of IEEE CVPR 2008, pp. 1–8 (2008)
12. Lassoued, I., Zagrouba, E., Chahir, Y.: An efficient approach for video action classification based on 3D Zernike moments. In: Park, J.J., Yang, L.T., Lee, C. (eds.) FutureTech 2011. CCIS, vol. 185, pp. 196–205. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22309-9_24
13. Newton, A.R., Hse, H.: Sketched symbol recognition using Zernike moments. In: Proceedings of the 17th ICPR, vol. 1, pp. 367–370 (2004)
14. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: Proceedings of CVPR 2008 (2008)
15. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM ICM, pp. 357–360 (2007)

16. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in NIPS 27, pp. 568–576 (2014)
17. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of the IEEE ICCV 2003, vol. 2, pp. 1470–1477 (2003)
18. Tsolakidis, D.G., Kosmopoulos, D.I., Papadourakis, G.: Plant leaf recognition using Zernike moments and histogram of oriented gradients. In: Likas, A., Blekas, K., Kalles, D. (eds.) SETN 2014. LNCS (LNAI), vol. 8445, pp. 406–417. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07064-3_33
19. Wang, H., Klaser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: Proceedings of the IEEE CVPR 2011, pp. 3169–3176 (2011)
20. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_48