Space-Time Zernike Moments and Pyramid Kernel Descriptors for Action Classification

Luca Costantini[2], Lorenzo Seidenari[1], Giuseppe Serra[1], Licia Capodiferro[2], and Alberto Del Bimbo[1]

[1] Media Integration and Communication Center, University of Florence, Italy
{delbimbo,seidenari,serra}@dsi.unifi.it
[2] Fondazione Ugo Bordoni, Roma, Italy
{lcostantini,lcapodiferro}@fub.it

Abstract. Action recognition in videos is a relevant and challenging task of automatic semantic video analysis. Most successful approaches exploit local space-time descriptors. These descriptors are usually carefully engineered in order to obtain feature invariance to photometric and geometric variations. The main drawback of space-time descriptors is high dimensionality and efficiency. In this paper we propose a novel descriptor based on 3D Zernike moments computed for space-time patches. Moments are by construction not redundant and therefore optimal for compactness. Given the hierarchical structure of our descriptor we propose a novel similarity procedure that exploits this structure comparing features as pyramids. The approach is tested on a public dataset and compared with state-of-the art descriptors.

Keywords: video annotation, action classification, Zernike moments.

1 Introduction and Related Works

Human behavior recognition is a challenging computer vision task that have recently attracted wide research effort; this is mainly due to the need of automatic semantic analysis of video data in several application fields such as intelligent video-surveillance systems and digital libraries. In video surveillance it is often the case that human operators are simply not able to attentively observe a large amount of screens in parallel; moreover in forensics, retrieval of video footage containing well defined human actions is invaluable.

Several techniques have been developed in the recent years mainly based on the use of local descriptions of the imagery. Following the success of SIFT [1] in object and scene recognition and classification [2], several space-time extensions of the local patch descriptors have been proposed. Similarly to local image features [3,4] space-time features are localized through a detection step and then computed on the extracted patches; videos are represented as a collection of descriptors. Space-time descriptors represent the appearance and the motion of a local region and are engineered in order to retain invariance to geometric and

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 199–208, 2011.

photometric transformations. Laptev *et al.* [5] defined a descriptor as a concatenation of histograms of oriented 2D gradients and histograms of optical flow. In order to reduce the computation burden an extension of SURF have been presented in [6]. Scovanner *et al.* [7] extended the SIFT to three-dimensional gradients normalizing 3D orientations bins by the respective solid angle in order to cope with the issue of the uneven quantization of solid angles in a sphere. To solve this issue Kläser *et al.* [8] proposed to exploit 3D pixel gradients developing a technique based on Platonic solids. Finally Ballan *et al.* [9] developed an efficient descriptor decorrelating the spatial and temporal components and creating separated histograms of 3D gradient orientations. However, all of these descriptors are extremely high-dimensional and often retain redundant information.

In the same time, researchers have exploited moments and invariant moments in pattern recognition [10]. Moments are scalar quantities used to characterize a function and to capture its significant features and they have been widely used for hundreds of years in statistics for description of the shape of a probability density function. Moments and in particular Zernike moments are a common choice in shape representation [11]. Zernike moments have been also proposed in action recognition as holistic features in [12] to describe the human silhouettes.

Despite the fact that feature matching is an important step in the recognition process few works have analysed it. Lowe [1] showed that in order to retrieve meaningful patches it is necessary to look at the distances of the second nearest neighbour. More recently Bo *et al.* [13] provided a kernel view of the matching procedure between patches. Their work formulates the problem of similarity measurement between image patches as a definition of kernels between patches. Since these kernels are valid Mercer kernels it is straightforward to combine or plug them into kernelized algorithms.

In this paper we propose a new method for classification of human actions based on an extension of the Zernike moments to the spatio-temporal domain. Furthermore, we propose a kernel suitable for matching descriptors that can be hierarchically decomposed in order to obtain a multiple resolution representation. This kernel is inspired by multi-resolution matching of sets of features [14,15], but instead of matching sets of features we match single space-time patches at multiple resolutions. To the best of our knowledge 3D Zernike moments have never been used as local space-time features and the pyramid matching scheme has never been used to define kernels between single features but only to match sets of features. Experimental results on KTH dataset shows that our system presents a low computational time maintaining comparable performance with respect to the state-of-the-art. The rest of the paper is organized as follows. The generalization of the Zernike moments to the three dimensions is presented in the next section. The Pyramid Kernel Descriptors are introduced in Sect. 3. The techniques for action representation and classification are presented in Sect. 4. Experimental results on the standard KTH dataset are discussed in Sect. 5. Finally, conclusions are drawn in Sect. 6

2 Space-Time Zernike Moments

We first describe the formulation of the Zernike moments in two dimensions, and then introduce the generalization to the space-temporal domain. Let $\mathbf{x} = [x_1, x_2]$ be the Cartesian coordinates in the real plane $\mathbb{R}^2$. Zernike polynomials are a set of orthogonal functions within the unit disk composed by a radial profile R_{nm} and a harmonic angular profile $H_m(\vartheta)$ defined as follows

$$V_{nm}(\rho, \vartheta) = R_{nm}(\rho) \cdot H_m(\vartheta) \tag{1}$$

where $\rho = \sqrt{x_1^2 + x_2^2}$, $\vartheta = \tan^{-1}\left(\frac{x_2}{x_1}\right)$, $H_m(\vartheta) = e^{im\vartheta}$ and

$$R_{nm}(\rho) = \begin{cases} \sum\limits_{s=0}^{(n-|m|)/2} \frac{(-1)^s (n-s)! \rho^{n-2s}}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} & \text{for } n-|m| \text{ even} \\ 0 & \text{for } n-|m| \text{ odd} \end{cases}. \tag{2}$$

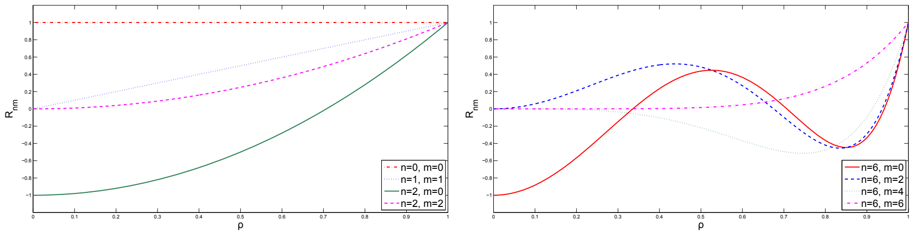
The index n is named “order” and is a non-negative integer, and m is called “repetition” and it is an integer such that $n - |m|$ is even and non-negative. In Fig. 1 some examples of the radial profile R_{nm} are shown. Both the Zernike polynomials and the radial profile $R_{nm}(\rho)$ satisfy the orthogonal condition

$$\int_0^{2\pi}\int_0^1 V_{nm}^*(\rho, \vartheta) V_{n'm'}(\rho, \vartheta) \rho d\rho d\vartheta = \frac{\pi}{n+1}\delta_{nn'}\delta_{mm'} \tag{3}$$

and

$$\int_0^1 R_{nm}(\rho) R_{n'm'}(\rho) \rho d\rho = \frac{1}{2(n+1)}\delta_{nn'}\delta_{mm'} \tag{4}$$

where δ indicates the Kronecker delta. Zernike polynomials are widely used to compute the Zernike moments [16,17].



(a) (b)

Fig. 1. a) Radial profile up to the 2^{nd} order; b) Radial profile for the 6^{nd} order

Let $f(\mathbf{x})$ be any continuous function, the Zernike moments are

$$A_{nm}(\mathbf{x_0}) = \frac{n+1}{\pi}\int\int_{\|\mathbf{x}-\mathbf{x_0}\|\leq 1} f(\mathbf{x}) V_{nm}^*(\mathbf{x}-\mathbf{x_0}) dx_1 dx_2 \tag{5}$$

where $\mathbf{x_0}$ denotes the point where the unit disk is centered. In this work we are interested in the computation of the Zernike moments for functions as $f : \mathbb{R}^3 \mapsto \mathbb{R}$ where the third dimension is the time. To get the 3D Zernike polynomials [18,19], the harmonic angular profile is substituted by the spherical harmonic functions

$$Y_m^l\left(\vartheta,\varphi\right) = N_m^l P_m^l\left(\cos\vartheta\right) e^{il\varphi} \tag{6}$$

where P_m^l denotes the Legendre function and N_m^l is a normalization factor

$$N_m^l = \sqrt{\frac{2m+1}{4\pi}\frac{(m-l)!}{(m+l)!}}. \tag{7}$$

The spherical harmonic functions up to the 3^{rd} order are shown in Fig. 2. In this case, given an order n, we use only the values of $m \geq 0$, and the index l is an integer such as $-m \leq l \leq m$. Then, the 3D Zernike polynomials are defined in spherical coordinates as follows

$$V_{nm}^l\left(\rho,\vartheta,\varphi\right) = R_{nm}\left(\rho\right)\cdot Y_m^l\left(\vartheta,\varphi\right) \tag{8}$$

and they satisfy the orthogonal condition within the unit sphere

$$\int_0^1\int_0^\pi\int_0^{2\pi}\left[V_{nm}^l\left(\rho,\vartheta,\varphi\right)\right]^* V_{n'm'}^{l'}\left(\rho,\vartheta,\varphi\right)\sin\left(\vartheta\right)d\vartheta d\varphi d\rho = \delta_{nn'}\delta_{mm'}\delta^{ll'}. \tag{9}$$

Let $\boldsymbol{\xi} = [\mathbf{x}, t]$ be the generic point in the real plane $\mathbb{R}^2$ at the time t, the 3D Zernike moments are

$$A_{nm}^l\left(\boldsymbol{\xi_0}\right) = \frac{3}{4\pi}\int_{\|\boldsymbol{\xi}-\boldsymbol{\xi_0}\leq 1\|} f\left(\boldsymbol{\xi}\right)\left[V_{nm}^l\left(\frac{\boldsymbol{\xi}-\boldsymbol{\xi_0}}{\boldsymbol{\sigma}}\right)\right]^* d\boldsymbol{\xi} \tag{10}$$

where $\boldsymbol{\xi_0}$ is the point where the unit sphere is centered, and $\boldsymbol{\sigma}$ tunes the size in pixel of the unit sphere for each coordinate. This $\boldsymbol{\sigma}$ is necessary because the patches, that we need to describe by using the 3D Zernike moments, can have different sizes in space and time. We use these space-time Zernike moments as descriptors for the local patches. The orthogonal condition (see Eq. 9) ensures that there is no redundant information in the descriptor allowing to have a compact representation of the local feature. Fig. 3 shows that we can obtain a rough but representative reconstruction of space-time cuboid from the 3D Zernike moments. In particular, we exploit the phase of these complex moments since from preliminary experiments proved to be more effective.

3 Pyramid Kernel Descriptors

We introduce a descriptor matching kernel inspired by multi-resolution matching of sets of features[15,14]; Grauman and Darrel [15] proposed the Pyramid

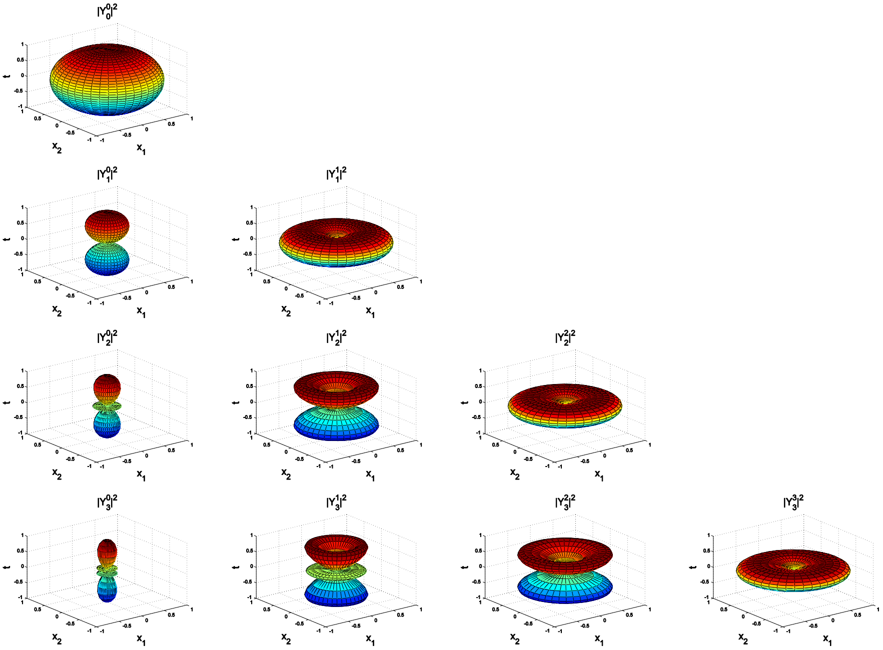


Fig. 2. Spherical harmonic functions up to the 3^{rd} order

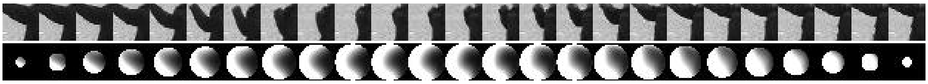


Fig. 3. Frames of a cuboid (top). Reconstructed cuboid from complex 3D Zernike moments up to the 6^{th} order (bottom).

Matching kernel to find an approximate correspondence between two sets of features points. Informally, their method takes a weighted sum of the number of matches that occur at each level of resolution, which are defined by placing a sequence of increasingly coarser grids over the features space. At any resolution, two feature points match if they fall into the same cell of the grid; number of matches computed at finer resolution are weighted more than those at coarser resolution. Later, Lazebnik *et al.* [14] introduced the Spatial Pyramid Matching kernel that work by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-regions.

Differently from these approaches our idea is to adapt the pyramid scheme for computing the similarity between two descriptor points. This allows to compute the similarity between two descriptors at multiple resolutions, exploiting a more distinctive representation when available and discarding it when at higher resolutions becomes noisy. We call our proposed approach "Pyramid Kernel Descriptors" because feature points are matched considering the descriptors as a multi-resolution set.

We consider a set of space-time interest points $X = \{\xi_1, \dots \xi_s\}$ and their descriptors $D = \{d_1, \dots, d_s\}$, where each descriptor can be organized in p sets $\{s^1, \dots, s^p\}$ hierarchically ordered. The pyramid kernel between d_i and d_j is defined as a weighted sum of the similarities of sets found at each level of the pyramid:

$$K(d_i, d_j) = \sum_{k=0}^{p} w_k k_c(s_i^k, s_j^k) \tag{11}$$

where w_k is the weight and $k_c(s_i^k, s_j^k)$ is a kernel to compute similarity between s_i^k and s_j^k. The similarity found at each level in the pyramid is weighted according to the description resolution: similarities made at a finer resolution, where features are most distinct, are weighted more than those found at a coarser level. Thus, if the p sets are arranged in ascending order the weight at level k can be defined as $w_k = 2^{k-p}$. If k_c is a valid kernel, our proposed kernel is a valid Mercer kernel for the closure property of kernels since it is a weighted sum of valid kernels. As described in sect. 2, our description based on space-time Zernike moments have a pyramid structure defined by the orders. In fact, lower order moments describe low frequencies of each cuboid while higher order moments encode higher frequencies. We define s^k as the concatenation of the phases of the complex Zernike moments for the first k orders: $s^k = \big(\arg(A_{00}^0), \dots, \arg(A_{km}^l)\big)$, where m and l are set according to Sect. 2. We use a normalized scalar product: $k_c(s_i^k, s_j^k) = \frac{s_i^k \cdot s_j^k}{\|s_i^k\| \|s_j^k\|}$, as a kernel between s_i^k and s_j^k, which is a valid Mercer kernel. Note that we normalize the scalar product computed at each level in order to have comparable values in the final sum.

For example, if we use a two level pyramid kernel descriptor then $s_0 = \big(\arg(A_{00}^0)\big)$, $s_1 = \big(\arg(A_{00}^0), \arg(A_{11}^{-1}), \arg(A_{11}^0), \arg(A_{11}^1)\big)$ and the corresponding weights are $w_0 = 1$ and $w_1 = \frac{1}{2}$. The final kernel between two space-time Zernike descriptors d_i, d_j computed up to the n^{th} order is:

$$K(d_i, d_j) = \sum_{k=0}^{n} 2^{k-n} \frac{s_i^k \cdot s_j^k}{\|s_i^k\| \|s_j^k\|}. \tag{12}$$

4 Action Classification

We represent an action as a bag of space-time interest points detected by an adaptation of the detector proposed by Dollár *et al.* [20]. This detector applies two separate linear filters to spatial and temporal dimensions, respectively. The response function has the form:

$$R = \Big(I * g_\sigma * h_{ev}\Big)^2 + \Big(I * g_\sigma * h_{od}\Big)^2 \tag{13}$$

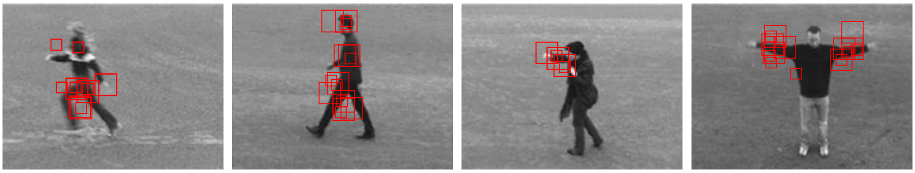


Fig. 4. Examples of space-time interest points extracted at multiple scales for different actions. Clips are taken from the KTH dataset: running, walking, boxing and hand-waving.

where $I(x, y, t)$ is a sequence of images over time, $g_\sigma(x, y)$ is the spatial Gaussian filter with kernel σ, h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied along the time dimension. They are defined as $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$, where $\omega = 4/\tau$. The interest points are detected at locations where the response is locally maximum typically corresponding to the temporal intensity changes. In order to cope with spatial and temporal scale variations we extract features at multiple scales. Usually these locations correspond to human body limbs involved in the execution of an action as can be seen in Fig. 4.

Each point is described using Space-time Zernike moments and then a nearest-neighbor classifier based on the concept of instance-to-class similarity [21] is used for action categorization. We choose not to employ descriptor codebooks (as in bag-of-words approaches) in order to better evaluate the effectiveness of our descriptor alone.

The instance-to-class nearest-neighbor classifier estimates the class posterior probability given the query video clip with a non-parametric density estimation based on local Parzen windows centered on descriptors belonging to the class. In [21] authors have shown that formulations based on more than one nearest neighbor per query descriptor do not significantly outperforms the simpler 1-NN formulation. Given this evidence, the implementation of this simple but effective classifier boils down to obtaining the most similar descriptor from the database for each feature extracted in a query clip (generally based on Euclidean distance between descriptors) and accumulating a vote for the class to which the database descriptor belongs to. The class with more votes is associated to the query clip. Instead of using Euclidean distance, we use our pyramid kernel descriptors (Sect. 3) to select the most similar descriptors which have, for each feature, the maximum kernel values.

5 Experimental Results

We tested our approach on the KTH action dataset containing six actions (walking, running, jogging, hand-clapping, hand-waving, boxing) performed several times by 25 actors under four different scenarios of illumination, appearance and scale change. The dataset contains 2391 video sequences with resolution of 160×120 pixel.

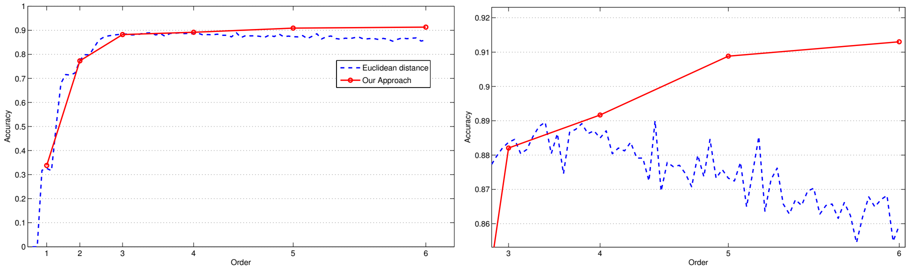
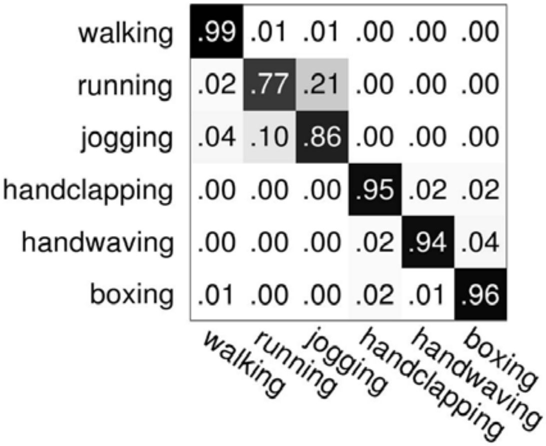


Fig. 5. Comparison of the two similarity techniques; right) detail showing the effect of pyramid matching descriptors on high order moments



	walking	running	jogging	handclapping	handwaving	boxing
walking	.99	.01	.01	.00	.00	.00
running	.02	.77	.21	.00	.00	.00
jogging	.04	.10	.86	.00	.00	.00
handclapping	.00	.00	.00	.95	.02	.02
handwaving	.00	.00	.00	.02	.94	.04
boxing	.01	.00	.00	.02	.01	.96

Fig. 6. Confusion matrix for the KTH dataset

We used a leave-one-out procedure specifically we used 24 actors' clips as a training set and the remaining actor's clips as a test set. Performance is presented as the average accuracy of 25 runs, each with a different person. First we tested our descriptor using the nearest-neighbor classifier based on the Euclidean distance and increasing the amount of moments (see Fig. 5). With this approach the use of high order moments degrades the performance of the classifier. This is due to the fact that the high order filters response in small scale cuboids is mostly noisy. Then we used our pyramid similarity kernel increasing the levels of detail. As discussed in Sect. 3 levels with higher order moments are weighted more than levels with lower order moments. We can see that in this case we can exploit the higher details captured by high order moments without degrading the overall classifier performance.

The confusion matrix reported in Fig. 6 shows that as expected jogging and running are the most difficult actions to discriminate while for all other classes results are quite satisfying.

Table 1. Descriptor complexity comparison together with accuracy

Method	Size	Computation time	Accuracy
Pyramid Zernike 3D	84	0.0300 s	91.30%
Gradient + PCA[20]	100	0.0060 s	81.17%
3D SIFT[7]	640	0.8210 s	82.60%
Ext Grad LBP-TOP + PCA[22]	100	0.1000 s	91.25%
3DGrad[9]	432	0.0400 s	90.38%
HOG-HOF[3][5]	162	0.0300 s	91.80%
HOG3D[3][8]	380	0.0020 s	91.40%
SURF3D[3][6]	384	0.0005 s	84.26%

In Tab. 1 we compare our descriptor with the state-of-the-art on KTH dataset respect to the computation time, storage needs and accuracy. Computation time is measured on our machine when the code was available while it is reported from the original publication if not. The accuracy is reported from the experiments reported in the original publication. We can see that Pyramid Zernike 3D descriptors are the smallest in terms of storage and are fast as other non-trivial implementations and C/C++ implementations; note that Gradient PCA is a simple concatenation of pixel gradient values and projection on principal components. Our descriptor is implemented without any optimization in MATLAB.

6 Conclusions

In this paper we have presented a method for action classification based on a new compact descriptor for spatio-temporal interest points. We introduce a new kernel suitable for matching descriptors that can be decomposed in multi-resolution sets. The approach was validated on the KTH dataset, showing results that have a low spatial and temporal computational complexity with comparable performance with the state-of-the-art. Our future work will deal with evaluation on more realistic datasets.

Acknowledgments. This work is partially supported by the EU IST euTV (contract FP7-226248).

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
2. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. of ICCV (2003)
3. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision 65(1-2) (2005)

[3] c++ implementation.

4. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10) (2005)
5. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. of CVPR (2008)
6. 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. ECCV, pp. 650–663. Springer, Heidelberg (2008)
7. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proc. of ACM Multimedia (2007)
8. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: Proc. of BMVC (2008)
9. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. In: Proc. of ICIP (2009)
10. Flusser, J., Zitova, B., Suk, T.: Moments and Moment Invariants in Pattern Recognition. Wiley Publishing, Chichester (2009)
11. Li, S., Lee, M.C., Pun, C.M.: Complex zernike moments features for shape-based image retrieval. IEEE Transactions on Systems, Man, and Cybernetics (2009)
12. Sun, X., Chen, M., Hauptmann, A.: Action recognition via local descriptors and holistic features. In: Proc. of Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB) (2009)
13. Bo, L., Ren, X., Fox, D.: Kernel descriptors for visual recognition. In: Advances in Neural Information Processing Systems (2010)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of CVPR (2006)
15. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: Proc. of ICCV (2005)
16. Neri, A., Carli, M., Palma, V., Costantini, L.: Image search based on quadtree zernike decomposition. Journal of Electronic Imaging 19(4) (2010)
17. Li, S., Lee, M.C., Pun, C.M.: Complex zernike moments features for shape-based image retrieval. IEEE Transactions on Systems, Man, and Cybernetics 39(1) (2009)
18. Canterakis, N.: 3d zernike moments and zernike affine invariants for 3d image analysis and recognition. In: Proc. of Conference on Image Analysis (1999)
19. Novotni, M., Klein, R.: Shape retrieval using 3d zernike descriptors. Computer-Aided Design 36(11), 1047–1062 (2004)
20. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proc. of VSPETS (2005)
21. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proc of. CVPR (2008)
22. Mattivi, R., Shao, L.: Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 740–747. Springer, Heidelberg (2009)