Effective Codebooks for Human Action Representation and Classification in Unconstrained Videos

Lamberto Ballan, *Member, IEEE*, Marco Bertini, *Member, IEEE*, Alberto Del Bimbo, *Member, IEEE*,
Lorenzo Seidenari, *Student Member, IEEE*, and Giuseppe Serra

***Abstract*—Recognition and classification of human actions for annotation of unconstrained video sequences has proven to be challenging because of the variations in the environment, appearance of actors, modalities in which the same action is performed by different persons, speed and duration, and points of view from which the event is observed. This variability reflects in the difficulty of defining effective descriptors and deriving appropriate and effective codebooks for action categorization. In this paper, we propose a novel and effective solution to classify human actions in unconstrained videos. It improves on previous contributions through the definition of a novel local descriptor that uses image gradient and optic flow to respectively model the appearance and motion of human actions at interest point regions. In the formation of the codebook, we employ radius-based clustering with soft assignment in order to create a rich vocabulary that may account for the high variability of human actions. We show that our solution scores very good performance with no need of parameter tuning. We also show that a strong reduction of computation time can be obtained by applying codebook size reduction with Deep Belief Networks with little loss of accuracy.**

***Index Terms*—Human action categorization, spatio-temporal local descriptors, visual codebooks.**

I. Introduction and Related Work

With the continuous growth of video production and archiving, the need for automatic annotation tools that enable effective retrieval by content has accordingly gained increasing importance. In particular, action recognition is a very active research topic with many important applications such as human–computer interaction, video indexing, and video-surveillance. Existing approaches for human action recognition can be classified as using holistic or part-based information [1], [2]. Most of the holistic-based methods usually perform better in a controlled environment and are computationally expensive due to the requirement of preprocessing the input data. Moreover, these representations can be influenced by motions of multiple objects, variations in the background and occlusions. Instead, part-based representations that exploit interest point detectors combined with robust feature descriptors have been used very successfully for object and scene classification tasks in images [3], [4]. As a result, nowadays most video annotation solutions have exploited the bag-of-features approach to generate textual labels that represent the categories of the main and easiest to detect entities (such as objects and persons) in the video sequence [5], [6].

The definition of effective descriptors that are able to capture both spatial and temporal features has opened the possibility of recognizing dynamic concepts in video sequences. In particular, interesting results have been obtained in the definition of solutions to automatically recognize human body movements, which usually represent a relevant part of video content [7]–[10]. However, the recognition and classification of such dynamic concepts for annotation of generic video sequences has proven to be very challenging because of the very many variations in environment, people and occurrences that may be observed. These can be caused by cluttered or moving background, camera motion, and illumination changes; people may have different size, shape, and posture appearance; semantically equivalent actions can manifest differently or partially, due to speed, duration, or self-occlusions; the same action can be performed in different modes by different persons. This great variability on the one hand reflects in the difficulty of defining effective descriptors and on the other makes it hard to obtain a visual representation that may describe such dynamic concepts appropriately and efficiently.

A. Effective Spatio-Temporal Descriptors

Holistic descriptors of body movements have been proposed by a few authors. Among the most notable solutions, Bobick *et al.* [11] proposed motion history images and their low-order moments to encode short spans of motion. For each frame of the input video, the motion history image is a gray scale image that records the location of motion; recent motion results into high intensity values whereas older motion produces lower intensities. Efros *et al.* [12] created stabilized spatio-temporal volumes for each action video segment and extracted a smoothed dense optic flow field for each volume. They have proved that this representation is particularly suited for distant objects, where the detailed information of the appearance is not available. Yilmaz and Shah [13] used a spatio-temporal volume, built stacking object regions; descriptors encoding direction, speed, and local shape of the resulting 3-D surface were generated by

Manuscript received August 18, 2011; revised January 02, 2012 and March 02, 2012; accepted March 11, 2012. Date of publication March 19, 2012; date of current version July 13, 2012. This work was supported in part by the EU "euTV" Project under Contract FP7-226248. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pascal Frossard.

The authors are with the Media Integration and Communication Center, University of Florence, 50139 Florence, Italy (e-mail: ballan@dsi.unifi.it; bertini@dsi.unifi.it; delbimbo@dsi.unifi.it; seidenari@dsi.unifi.it; serra@dsi.unifi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2191268

measuring local differential geometrical properties. Gorelick *et al.* [14] analyzed 3-D shapes induced by the silhouettes and exploited the solution to the Poisson equation to extract features, such as shape structure and orientation. Global descriptors that jointly encode shape and motion were suggested in Lin *et al.* [15]; Wang *et al.* [16] exploited global histograms of optic flow together with hidden conditional random fields. Although encoding much of the visual information, these solutions have shown to be highly sensitive to occlusions, noise and change in viewpoint. Most of them have also proven to be computationally expensive due to the fact that some preprocessing of the input data is needed, such as background subtraction, segmentation, and object tracking. All of these aspects make these solutions only suited for representation of body movements in videos taken in controlled contexts.

Local descriptors have shown better performance and are in principle better suited for videos taken in both constrained and unconstrained contexts. They are less sensitive to partial occlusions and clutter and overcome some of the limitations of the holistic models, such as the need of background subtraction and target tracking. In this approach, local patches at spatio-temporal interest points are used to extract robust descriptors of local moving parts and the bag-of-features approach is employed to have distinctive representations of body movements. Laptev [17] and Dollár [18] approaches have been among the first solutions. Laptev [17], [19] proposed an extension to the Harris–Förstner corner detector for the spatio-temporal case; interesting parts were extracted from voxels surrounding local maxima of spatio-temporal corners, i.e., locations of videos that exhibit strong variations of intensity both in spatial and temporal directions. The extension of the scale-space theory to the temporal dimension permitted to define a method for automatic scale selection. Dollár *et al.* [18] proposed a different descriptor than Laptev's, by looking for locally periodic motion. While this method produces a denser sampling of the spatio-temporal volume, it does not provide automatic scale selection. Despite this, experimental results have shown that it improves with respect to [19].

Following these works, other authors have extended the definition of local interest point detectors and descriptors to incorporate time or combined static local features with other descriptors so to model the temporal evolution of local patches. Sun *et al.* [20] have fused spatio-temporal SIFT points with holistic features based on Zernike moments. In [21], Willems *et al.* extended SURF feature to time and defined a new scale-invariant spatio-temporal detector and descriptor that showed high efficiency. Scovanner *et al.* [22], have proposed to use grouping of 3-D SIFT, based on co-occurrence, to represent actions. Kläser *et al.* [23] have proposed a descriptor based on histograms of oriented 3-D gradients, quantized using platonic solids. Gao *et al.* [24] presented MoSIFT, an approach that extend the SIFT algorithm to find visually distinctive elements in the spatial domain. It detects spatio-temporal points with a high amount of optical flow around the distinctive points motion constraints. More recently, Laptev *et al.* [25] proposed a structural representation based on dense temporal and spatial scale sampling, inspired by the spatial pyramid approach of [26] with interesting classification results in generic video scenes. Kovashka *et al.* [27] extended this work by defining a hierarchy of discriminative neighborhoods instead of using spatio-temporal pyramids. Liu *et al.* [28] combined MSER and Harris–Affine [29] regions with Dollár's space-time features and used AdaBoost to classify YouTube videos. Shao *et al.* [30] applied transformation based techniques (i.e., discrete Fourier transform, discrete cosine transform, and discrete wavelet transform) on the local patches and used the transformed coefficients as descriptors. Yu *et al.* [31] presented good results using the Dollar's descriptor and random forest-based template matching. Niebles *et al.* [32] trained an unsupervised probabilistic topic model using the same spatio-temporal features, while Cao *et al.* [33] suggested to perform model adaptation in order to reduce the amount of labeled data needed to detect actions in videos of uncontrolled scenes. Comparative evaluations of the performance of the most notable approaches were recently reported by Wang *et al.* [34] and Shao *et al.* [1].

B. Suitable Visual Codebooks

According to the bag-of-features model actions are defined as sets of codewords obtained from the clustering of local spatio-temporal descriptors. Most of the methods have used the k-means algorithm for clustering because of its simplicity and speed of convergence [3], [32], [35], [36]. However, both the intrinsic weakness of k-means to outliers and the need of some empirical pre-evaluation of the number of clusters hardly fit with the nature of the problem at hand. Moreover, with k-means, the fact that cluster centers are selected almost exclusively around the most dense regions in the descriptor space results in ineffective codewords of action primitives. To overcome the limitations of the basic approach, Liu *et al.* [37] suggested a method to automatically find the optimal number of visual word clusters through maximization of mutual information (MMI) between words and actions. MMI clustering is used after k-means to discover a compact representation from the initial codebook of words. They showed some performance improvement. Recently, Kong *et al.* [38] have proposed a framework that unifies reduction of descriptor dimensionality and codebook creation, to learn compact codebooks for action recognition optimizing class separability. Differently, Uemura and Mikolajczyk [39] explored the idea of using a large number of features represented in many vocabulary trees instead of a single flat vocabulary. Yao *et al.* [40] recently proposed a similar framework using a training procedure based on a Hough voting forest. Both of these methods require higher efforts in the training phase.

C. Our Contribution

In this paper, we propose a novel and effective solution to classify human actions in unconstrained videos. It improves on previous contributions in the literature through the definition of a novel local descriptor and the adoption of a more effective solution for the codebook formation. We use image gradient and optic flow to respectively model the appearance and motion of human actions at regions in the neighborhood of local interest points and consider multiple spatial and temporal scales. These two descriptors are used in combination to model local features

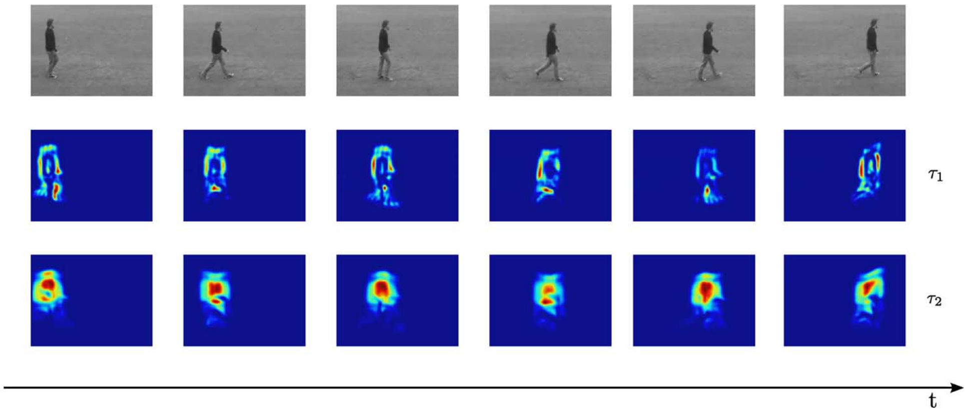


Fig. 1. Response of the spatio-temporal interest point detector at two temporal scales $\tau_1 < \tau_2$ (low response in blue, high response in red); first row: original video frames; second row: detector response at temporal scale τ_1 (mostly due to motion of human limbs); third row: detector response temporal scale τ_2 (mostly due to motion of the human torso).

of human actions and activities. Unlike similar related works [22], [23], no parameter tuning is required.

In the formation of the codebook we recognize that the clusters of spatio-temporal descriptors should be both in a sufficiently large number and sufficiently distinguished from each other so to represent the augmented variability of dynamic content with respect to the static case. To this end, radius-based clustering [41] with soft assignment has been used. In fact, with radius-based clustering, cluster centers are allocated at the modes corresponding to the maximal density regions, thus resulting in statistics of the codewords that better fit with the variability of human actions with respect to k-means clustering. Experiments carried on standard datasets show that the approach followed outperforms the current state of the art methods. To avoid too large codebooks, we performed codebook compression with Deep Belief Networks. The solution proposed shows good accuracy even with very small codebooks. Finally, we provide several experiments on the Hollywood2 dataset [42] and on a new surveillance dataset (MICC-Surveillance), to demonstrate the effectiveness and generality of our method for action recognition in unconstrained video domains. A preliminary version of this manuscript appeared at ICCV-VOEC 2009 [43], presenting our spatio-temporal features and clustering approach. This paper presents a novel codebook reduction technique, greatly extended experiments (also adding two novel challenging datasets), and more detailed information on the proposed method.

The remainder of this paper is organized as follows. The descriptor is presented in Section II. Action representation and categorization is presented in Section III. The experimental results, with an extensive comparison with the state-of-the-art approaches, are hence discussed in Section IV. Here, we also included experiments on unconstrained videos to demonstrate the effectiveness of the approach also in this case. Conclusions are drawn in Section V.

II. Spatio-temporal Local Descriptors of Appearance and Motion

Spatio-temporal interest points are detected at video local maxima of the Dollár's detector [18] applied over a set of spatial and temporal scales. Using multiple scales is fundamental to capture the essence of human activity. To this end, linear filters are separately applied to the spatial and temporal dimension: on the one hand, the spatial scale permits to detect visual features of high and low detail; on the other, the temporal scale allows to detect *action primitives* at different temporal resolutions. The filter response function is defined as

$$R = (I * g_\sigma * h_{ev})^2 + (I * g_\sigma * h_{od})^2 \quad (1)$$

where $I(x, y, t)$ is the image sequence, $g_\sigma(x, y)$ is a spatial Gaussian filter with scale σ, h_{ev} and h_{od} are a quadrature pair of 1-D Gabor filters that provide a strong response to temporal intensity changes for periodic motion patterns, respectively defined as

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2} \quad (2)$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2} \quad (3)$$

where $\omega = 4/\tau$. In the experiments, we used $\sigma = \{2, 4\}$ as spatial scales and $\tau = \{2, 4\}$ as temporal scales. Fig. 1 shows an example of temporal scaling of human body parts activity during walking: the torso has a high response at a high temporal scale, while limbs respond at the lower scale.

Three-dimensional regions of size proportional to the detector scale (6×) are considered at each spatio-temporal interest point, and divided into equally sized subregions (three for each spatial dimensions along the x and y axes, and two for the temporal dimension t), as shown in Fig. 2.

For each subregion, image gradients on x, y, and t are computed as

$$G_x = I(x+1, y, t) - I(x-1, y, t) \quad (4)$$

$$G_y = I(x, y+1, t) - I(x, y-1, t) \quad (5)$$

$$G_t = I(x, y, t+1) - I(x, y, t-1) \quad (6)$$

and the optic flow with relative apparent velocity V_x, V_y is estimated according to [44].

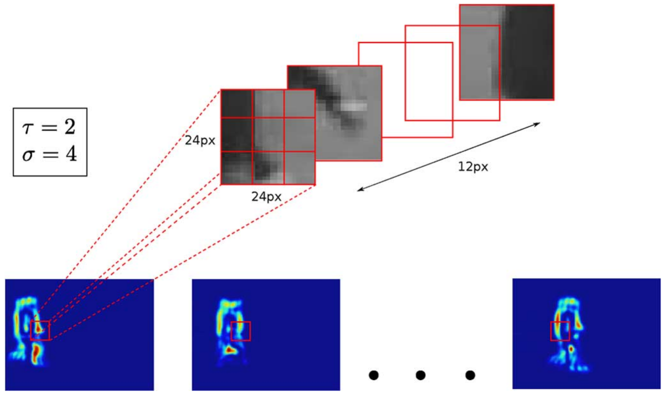


Fig. 2. Three-dimensional region at the spatio-temporal interest point corresponding to a swinging arm.

Orientations of gradients and optical flow are computed for each pixel as

$$\phi = \tan^{-1}\left(\frac{G_t}{\sqrt{G_x^2 + G_y^2}}\right) \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \tag{7}$$

$$\theta = \tan^{-1}\left(\frac{G_y}{G_x}\right) \in [-\pi, \pi] \tag{8}$$

$$\psi = \tan^{-1}\left(\frac{V_y}{V_x}\right) \in [-\pi, \pi] \tag{9}$$

where ϕ and the θ are quantized in four and eight bins, respectively.

The local descriptor obtained by concatenating ϕ and θ histograms (H3DGrad) has therefore size $3 \times 3 \times 2 \times (8+4) = 216$. There is no need to reorient the 3-D neighborhood, since rotational invariance, typically required in object detection and recognition, is not desirable in the action classification context. This approach is much simpler to compute than those proposed in [22] and [23]. In particular, in [22], the histogram is normalized by the solid angle value to avoid distortions due to the polar coordinate representation (instead of quantizing separately the two orientations as in our approach), moreover the size of the descriptor is 2048; in [23], the 3-D gradient vector is projected on the faces of a platonic solid. This latter approach requires additional parameter tuning to optimize the selection of the solid used for the histogram computation and whether to consider the orientations of its faces or not. Differently from [25], our 12-bin H3DGrad descriptor models the dynamic appearance of the 3-D region used for its computation, instead of being a 4-bin 2-D histogram cumulated over time. A comparison between our H3DGrad descriptor and the other HOG features (i.e., [22], [23], [25]) is reported in Table I, in terms of both accuracy and feature computation time.

The ψ is quantized in eight bins with an extra "no-motion" bin added to improve performance. The local descriptor of ψ (HOF) has size $3 \times 3 \times 2 \times (8+1) = 162$. Histograms of ϕ, θ, and ψ are respectively derived by weighting pixel contributions, respectively, with the gradient magnitude $M_G = \sqrt{G_x^2 + G_y^2 + G_t^2}$ (for ϕ and θ), and the optic flow magnitude $M_O = \sqrt{V_x^2 + V_y^2}$ (for ψ).

TABLE I
COMPARISON OF ACCURACY AND EFFICIENCY OF OUR H3DGRAD WITH OTHER GRADIENT-BASED DESCRIPTORS ON KTH AND WEIZMANN DATASETS. COMPUTATION TIME FOR A SINGLE DESCRIPTOR MEASURED ON A 2.66-GHz INTEL XEON WITH 12-GB RAM; H3DGRAD, [23] AND [25] ARE C++ IMPLEMENTATIONS WHILE [22] IS A MATLAB IMPLEMENTATION

Descriptor	**KTH**	**Weizmann**	**Time (ms)**
H3DGrad	90.38	92.30	1
Kläser *et al.* [23]	91.40	84.30	2
Laptev *et al.* [25]	81.60	-	12
Scovanner *et al.* [22]	-	82.60	419

In order to obtain an effective codebook for human actions, these two descriptors can be combined according to either early or late fusion. In the former case, the two descriptors are first concatenated and the combined descriptor is hence used for the definition of the human action codebook. In the latter, a codebook is obtained from each descriptor separately; then, the histograms of codewords are concatenated to form the representation (see Fig. 3).

Fig. 4 shows the classification accuracy measured with the KTH dataset, using codebooks based on (a) the H3DGrad descriptor, (b) HOF descriptor, and (c) early and (d) late fusion, with 4000 codewords. Each action is represented by a histogram H of codewords w obtained according to k-means clustering with hard assignment

$$H(w) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1, & \text{if } w = \arg\min\limits_{v \in V}(D(v, f_i))\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

where n is the number of the spatio-temporal features, f_i is the ith spatio-temporal feature, and $D(v, f_i)$ is the Euclidean distance between the codeword v of the vocabulary V and f_i.

We present in Table II the average accuracy obtained by H3DGrad and HOF, respectively, and by the early and late fusion. From the figures, it appears clearly that late fusion provides the best performance. This can be explained with the fact that H3DGrad and HOF descriptors have quite complementary roles [for example, the *boxing* action is better recognized when using H3DGrad descriptor while *hand-clapping* action is better recognized by HOF, as shown in Fig. 4(a) and (b)]. Late fusion improves recognition performance for all of the classes except one. A similar behavior was observed with the Weizmann dataset, although in this case the improvement was not so significant mainly due to the limited size and intra-class variability of the dataset (see Table II).

III. ACTION REPRESENTATION AND CLASSIFICATION

In order to improve with respect to k-means and to account for the high variability of human actions in terms of appearance or motion, we used radius-based clustering for codebook formation.

Fig. 5 shows the codeword frequency of radius-based clustering and k-means with hard quantization on the KTH dataset. It is interesting to note that, with k-means, most of the codewords have similar probability of occurrence, thus making it difficult to identify a set of words that simultaneously have high discrimination capability and good probability of occurrence. In

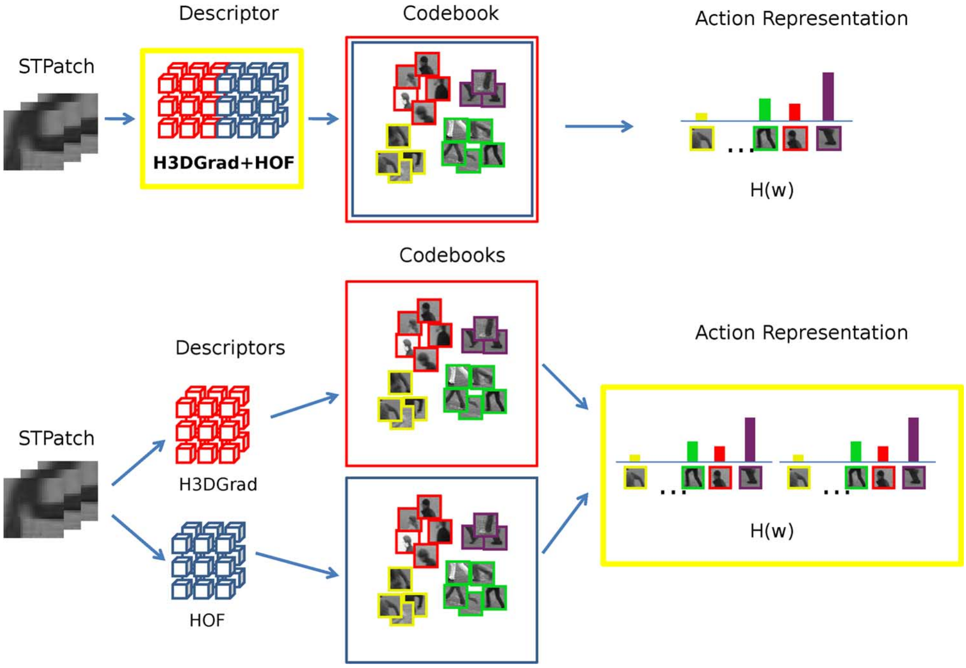


Fig. 3. Two fusion strategies: early fusion (at the descriptor level) and late fusion (at the codebook level).

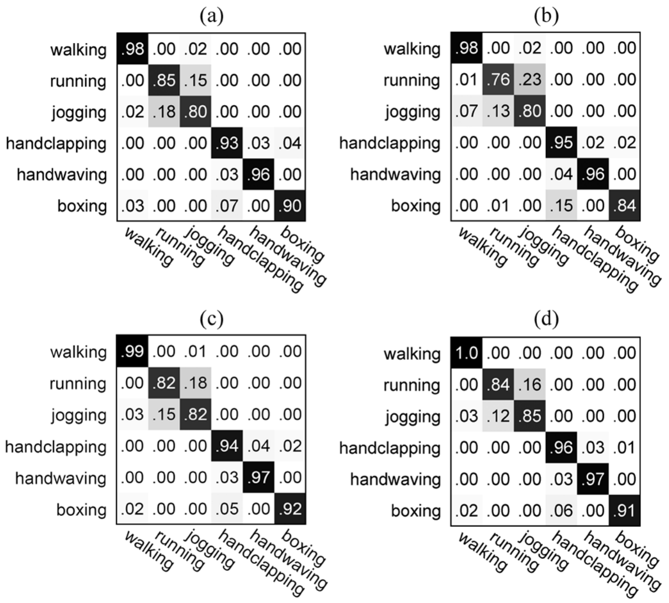


Fig. 4. Classification accuracy on the KTH dataset using k-means clustering, hard assignment, and different descriptors combination strategies (i.e., early or late fusion). (a) H3DGrad. (b) HOF. (c) H3DGrad+HOF (early). (d) H3DGrad+HOF (late).

TABLE II
AVERAGE CLASS ACCURACY OF OUR DESCRIPTORS, ALONE AND COMBINED, ON THE KTH AND WEIZMANN DATASETS

Descriptor	KTH	Weizmann
H3DGrad	90.38	92.30
HOF	88.04	89.74
H3DGrad + HOF (early fusion)	91.09	92.38
H3DGrad + HOF (late fusion)	**92.10**	**92.41**

contrast, radius-based shows a much less uniform frequency distribution. Interestingly, with radius-based clustering, the codeword distribution of the human action vocabulary is similar to the Zipf's law for textual corpuses. Therefore, it seems reasonable to assume that codewords at intermediate frequencies are the most informative also for human action classification and the best candidates for the formation of the codebook.

Due to the high dimensionality of the descriptor, codebooks for human actions usually have cluster centers that are spread in the feature space, so that two or more codewords are equally relevant for a feature point (codeword *uncertainty*); moreover, cluster centers are often too far from feature points so that they are not anymore representative (codeword *plausibility*). With radius-based clustering, codeword *uncertainty* is critical because it frequently happens that feature points are close to the codewords boundaries [46]. Instead, codeword *plausibility* is naturally relaxed due to the fact that clusters are more uniformly distributed in the feature space. To reduce the *uncertainty* in codeword assignment, we therefore performed radius-based clustering with soft assignment by Gaussian kernel density estimation smoothing. In this case, the histogram H is computed as

$$H(w) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_\sigma(w, f_i)}{\sum_{j=1}^{|V|} K_\sigma(v_j, f_i)} \qquad (11)$$

where K_σ is the Gaussian kernel, and $K_\sigma(\cdot,\cdot) = (1/\sqrt{2\pi}\sigma)e^{(-(d(\cdot,\cdot)^2/2\sigma^2))}$ where σ is the scale parameter tuned on the training set, and $d(\cdot,\cdot)$ is the Euclidean distance.

Fig. 6 compares the classification accuracy with codebooks obtained with k-means clustering with both hard and soft assignment and radius-based clustering with soft assignment, respectively for the KTH and Weizmann dataset. The plots have been obtained by progressively adding less frequent codewords to the codebooks (up to 4000 and 1000 codewords, respectively, for the two datasets). The performance of k-means is improved by the use of soft assignment. With a small number of words,

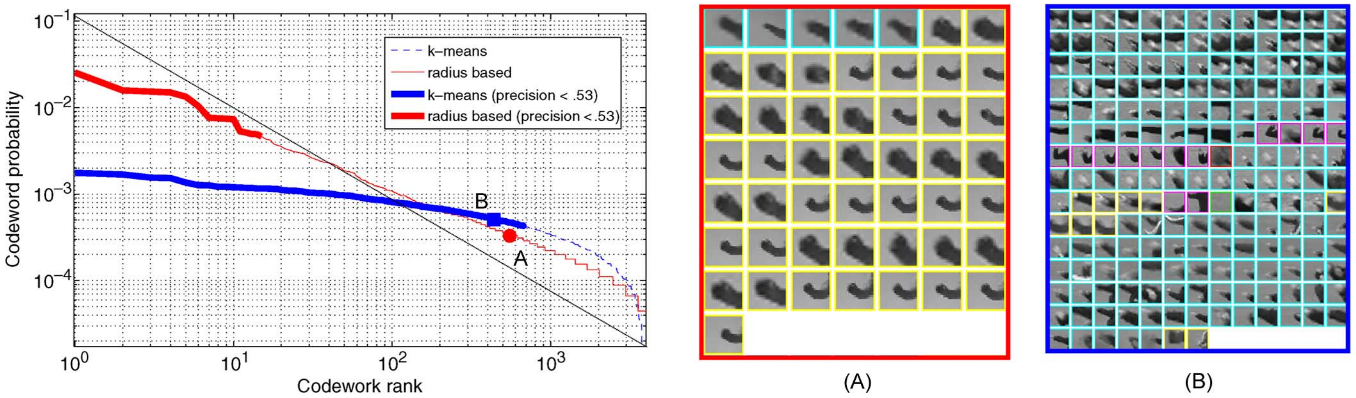


Fig. 5. Log-log plots of codeword frequency using k-means and radius-based clustering with hard assignment. Bold lines indicate regions where the average cluster precision [45] is below 0.53. The dotted diagonal line represents the Zipfian distribution. Two sample clusters are shown at near frequencies, respectively obtained with radius-based clustering (A) (most of the features in the cluster represent spatio-temporal patches of the same action) and with k-means (B) (features in the cluster represent patches of several actions). Patches of actions have different colors: *boxing* (cyan), *hand-waving* (magenta), *hand-clapping* (yellow), *running* (green), *walking* (red), and *jogging* (blue).

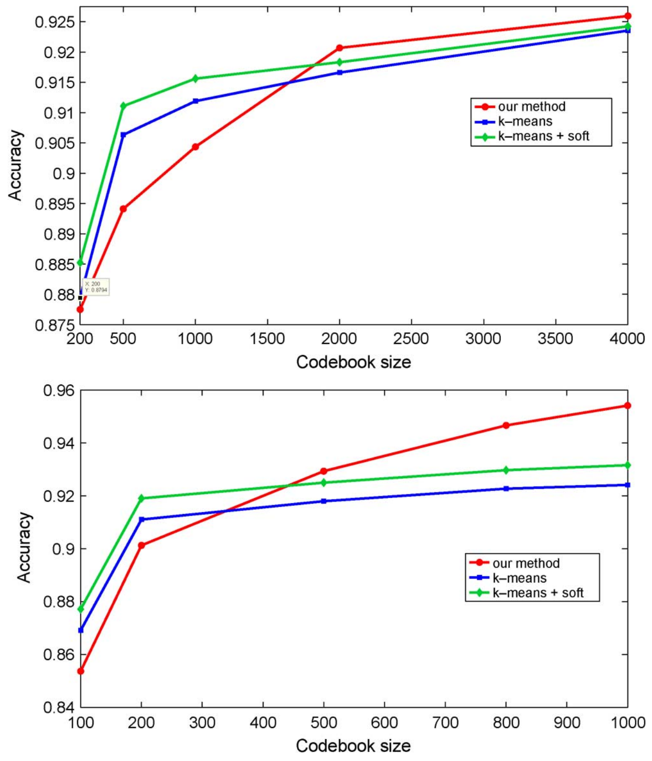
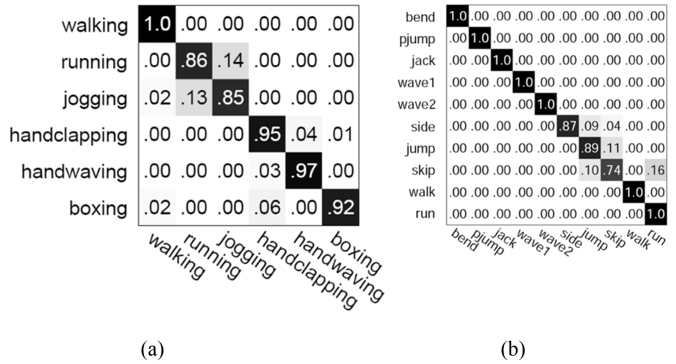


Fig. 6. Classification accuracy on KTH (top) and Weizmann (bottom) datasets with codebooks created with k-means with hard assignment, k-means with soft assignment and radius-based with soft assignment.



	walking	running	jogging	handclapping	handwaving	boxing
walking	1.0	.00	.00	.00	.00	.00
running	.00	.86	.14	.00	.00	.00
jogging	.02	.13	.85	.00	.00	.00
handclapping	.00	.00	.00	.95	.04	.01
handwaving	.00	.00	.00	.03	.97	.00
boxing	.02	.00	.00	.06	.00	.92

(a)

	bend	pjump	jack	wave1	wave2	side	jump	skip	walk	run
bend	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
pjump	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
jack	.00	.00	1.0	.00	.00	.00	.00	.00	.00	.00
wave1	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
wave2	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
side	.00	.00	.00	.00	.00	.87	.09	.04	.00	.00
jump	.00	.00	.00	.00	.00	.00	.89	.11	.00	.00
skip	.00	.00	.00	.00	.00	.00	.10	.74	.00	.16
walk	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00
run	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0

(b)

Fig. 7. Classification accuracy on (a) KTH and (b) Weizmann datasets using radius-based clustering with soft assignment.

radius-based clustering with soft assignment has lower performance than k-means due to the fact that the codewords used have higher frequency than those used by k-means (see Fig. 5). As the number of codewords in the codebook increases, radius-based clustering outperforms k-means, whether with hard or soft assignment. This reflects the fact that in this case radius-based clustering permits to have also sparse regions being represented in the codebook. In addition, soft assignment helps to reduce *uncertainty* in the dense regions. Fig. 7 shows the confusion matrix for different human actions on KTH and Weizmann datasets with radius-based soft assignment. The average accuracy is 92.66% and 95.41%, respectively, for the two datasets.

IV. Experimental Results

We have assessed our approach for categorization of human actions in different conditions. Particularly, it has been tested on the KTH and Weizmann datasets that show staged actions performed by an individual in a constrained noncluttered environment. Moreover, in order to have a more complete assessment of the performance of the proposed solution even in real-world scenes with high variability and unconstrained videos, we also carried out experiments on the Hollywood2 and MICC-UNIFI Surveillance datasets. This latter, made publicly available at www.openvisor.org [47], includes real-world video surveillance sequences containing actions performed by individuals with cluttering and varying filming conditions. Experiments were performed using nonlinear SVMs with the χ^2 kernel [4].

A. Experiments on KTH and Weizmann Datasets

The KTH dataset, currently the most common dataset used for the evaluations of action recognition methods [34], contains 2391 short video sequences showing six basic actions: *walking, running, jogging, hand-clapping, hand-waving, boxing*. They

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY WITH SOME STATE-OF-THE-ART METHODS ON KTH AND WEIZMANN DATASETS

Method	**KTH**	**Weizmann**	**Features**	**Optimizations**
Our method	**92.66**	**95.41**	H3DGrad + HOF	-
Yu *et al.* [31]	91.8		HoG + HOF	-
Wang *et al.* [34]	92.1	-	HOF	-
Gao *et al.* [24]	91.14	-	MoSIFT	-
Sun *et al.* [20]	89.8	90.3	2D SIFT + 3D SIFT + Zernike	-
Rapantzikos *et al.* [49]	88.3	-	PCA-Gradient	-
Laptev *et al.* [25]	91.8	-	HoG + HOF	codebook, sampling
Wong and Cipolla [50]	86.62	-	PCA-Gradient	-
Scovanner *et al.* [22]	-	82.6	3D SIFT	codebook
Liu *et al.* [48]	-	90.4	PCA-Gradient + Spin images	codebook
Kläser *et al.* [23]	91.4	84.3	3D HoG	descriptor
Willems *et al.* [21]	84.26	-	3D SURF	-
Schüldt *et al.* [19]	71.7	-	ST-Jets	-

are performed by 25 actors under four different scenarios with illumination, appearance, and scale changes. They have been filmed with a handheld camera at 160 × 120 pixel resolution. The Weizmann dataset contains 93 short video sequences showing nine different persons, each performing ten actions: *run, walk, skip, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallop-sideways, wave-two-hands, wave-one-hand*, and *bend*. They have been filmed with a fixed camera, at 180 × 144 pixel resolution, under the same lighting conditions.

Table III reports the average accuracy of our method in comparison with the most notable research results published in the literature. The performance figures reported are those published in their respective papers. For a fair comparison, our experiments have been performed with the setup suggested by the creators of the KTH and Weizmann datasets [14], [19], that has been used in [19]–[25], [31], [34], [48]–[50]. In particular, with the KTH dataset, SVM classifiers have been trained on sequences of 16 actors and performance was evaluated for the sequences of the remaining nine actors according to fivefold cross validation. With the Weizmann dataset SVM classifiers have been trained on the videos of eight actors and tested on the one remaining, following leave-one-out cross validation.

While showing the best performance, our solution has also the nice property that it does not require any adaptation to the context under observation. Instead other solutions require some tuning of the descriptor to the specific context. Namely, Laptev *et al.* [25] perform different spatio-temporal sampling of video frames and define a set of descriptors; hence, they represent each action with the best combination of sampling and descriptors. Kläser *et al.* [23] use a parameterized 3-D gradient descriptor. Parameter values are optimized for the dataset used. Liu *et al.* [48] use both local and global descriptors and select the best combination of them according to an optimization procedure. Scovanner *et al.* [22] optimize the codebook by associating co-occurrent visual words.

Other researchers have claimed higher performance on the KTH dataset: 93.17% Bregonzio *et al.* [51]; 94.2% Liu and Shah [37]; 93.43% Lin *et al.* [15]; and 95.83% Chen *et al.* [52]. However, these results were obtained with a leave-one-out cross-validation setting that uses more training data and therefore are not directly comparable. For the sake of fairness, they have not been included in Table III. An exhaustive list of the different experimental setups and results has been recently published by Gao *et al.* [24].

B. Experiments on MICC-UNIFI Surveillance Dataset

The MICC-UNIFI Surveillance dataset is composed by 175 real world video sequences of human actions with durations ranging from 3 to 20 s. The videos have been taken from wall mounted Sony SNC RZ30 cameras at 640 × 480 pixel resolution, in a parking lot. The scenes are captured from different viewpoints, at different degrees of zooming, with different shadowing and unpredictable occlusions, at different duration, speed, and illumination conditions. Eight subjects perform seven everyday actions: *walking, running, pickup object, enter car, exit car, handshake*, and *give object*. A few examples are shown in Fig. 8. We followed a repeated stratified random subsampling validation, using 80% of the videos of each class as training set. Experiments were performed using a 2000 codeword codebook. The confusion matrix of classification accuracy is reported in Fig. 9: the average accuracy is 86.28%. Most of the misclassifications observed with our method occurred with the *give object* and *handshake* actions. They are both characterized by a very fast motion pattern and small motion of the human limbs. Fig. 10 reports sample sequences of these actions with evidence of details. In Table IV, we report a comparison of our method with other codebook creation approaches (k-means with hard and soft assignment) and with other state-of-the-art descriptors that publicly make their implementation available: MoSIFT[1] [24] and Dollár *et al.*[2] [18]. The results show that the proposed method outperforms the other approaches and that the proposed codebook creation approach performs better than the typical k-means clustering whether with hard and soft assignment.

C. Experiments on Hollywood2 Dataset

The Hollywood2 dataset [42] is composed by sequences extracted from DVDs of 69 Hollywood movies, showing 12 different actions in realistic and challenging settings: *answer phone, drive car, eat, fight person, get out of car, handshake,*

[1] http://lastlaugh.inf.cs.cmu.edu/libscom/downloads.htm

[2] http://vision.ucsd.edu/%7epdollar/research.html

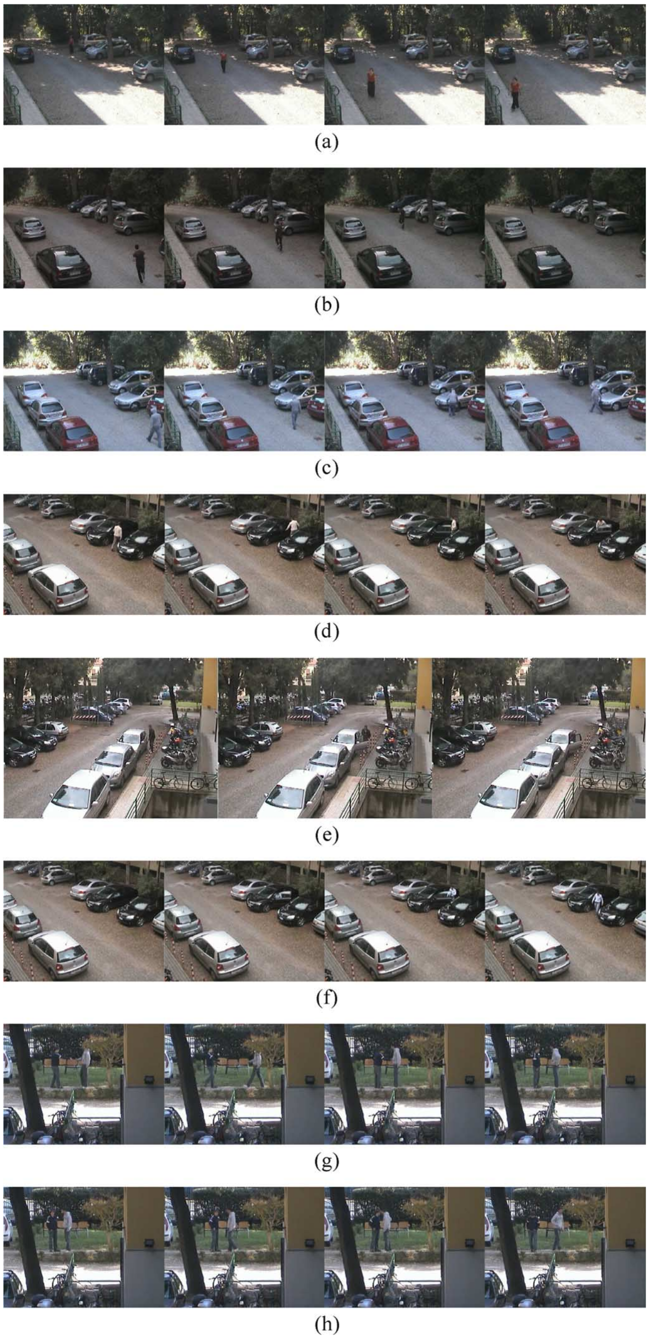
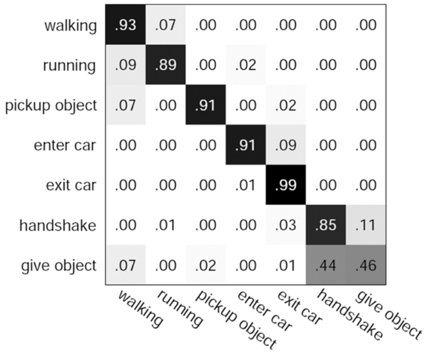


Fig. 8. Sample frames of sequences from the MICC-UNIFI Surveillance dataset. (a) Walking. (b) Running. (c) Pickup object. (d) Enter car. (e) Enter car (from a different view point). (f) Exit car. (g) Handshake. (h) Give object.



	walking	running	pickup object	enter car	exit car	handshake	give object
walking	.93	.07	.00	.00	.00	.00	.00
running	.09	.89	.00	.02	.00	.00	.00
pickup object	.07	.00	.91	.00	.02	.00	.00
enter car	.00	.00	.00	.91	.09	.00	.00
exit car	.00	.00	.00	.01	.99	.00	.00
handshake	.00	.01	.00	.00	.03	.85	.11
give object	.07	.00	.02	.00	.01	.44	.46

Fig. 9. Classification accuracy on the MICC-Surveillance dataset using radius-based clustering with soft assignment.

TABLE IV
COMPARISON OF CLASSIFICATION ACCURACY ON MICC-SURVEILLANCE DATASET WITH OUR METHOD, K-MEANS WITH SOFT ASSIGNMENT, K-MEANS WITH HARD ASSIGNMENT, AND WITH THE DESCRIPTORS PROPOSED IN [18] AND [24]

Method	**MICC-Surveillance**
Our method	**86.28**
k-means + soft	83.74
k-means	82.90
Dollár *et al.* [18]	72.50
MoSIFT [24]	75.88

TABLE V
COMPARISON OF PER-CLASS AP PERFORMANCE ON HOLLYWOOD2 DATASET WITH CODEBOOKS CREATED WITH OUR METHOD, K-MEANS WITH SOFT ASSIGNMENT, K-MEANS WITH HARD ASSIGNMENT AND WITH THE DETECTOR+DESCRIPTOR PROPOSED BY LAPTEV *et al.* [25]

Action	*k-means*	*k-means + soft*	*Our method*	Laptev *et al.* [25]
Answer phone	0.178	0.186	**0.195**	0.134
Drive car	0.864	**0.865**	0.863	0.861
Eat	0.552	0.564	0.564	**0.596**
Fight person	0.564	0.557	0.578	**0.643**
Get put of car	0.362	0.364	**0.362**	0.297
Handshake	0.142	0.143	0.167	**0.179**
Hug person	0.251	0.257	0.275	**0.345**
Kiss	0.494	**0.510**	0.503	0.467
Run	0.631	0.636	**0.659**	0.619
Sit down	0.483	0.493	**0.509**	0.505
Sit up	0.215	0.231	**0.227**	0.143
Stand up	0.511	0.513	**0.514**	0.485
mean AP	0.437	0.443	**0.451**	0.439

hug person, kiss, run, sit down, sit up, stand up. We performed our experiments with the same setup of [25], [34] using the "clean" training dataset, containing scenes that have been manually verified. This dataset is composed by 1707 sequences divided in training set (823) and test set (884), with different frame size and frame rate; train and test set videos have been selected from different movies. To be comparable with other experimental results the performance has been evaluated computing the average precision (AP) for each class and reporting also the mean AP over all classes. Codebooks have been created using 4000 codewords, as in [34]. We have compared our codebook creation approach with k-means clustering using both soft and hard assignments, and with an implementation of the method proposed in [25] using the provided descriptor and detector.[3] Results are reported in Table V, showing that the proposed method outperforms the other approaches in the majority of action classes and in terms of mean AP.

D. Reducing the Codebook Size

Large codebooks, although being able to exploit the most informative codewords as illustrated in Fig. 5, imply high time and space complexity. Reduction of codebook size with preservation of descriptive capability is therefore desirable. Linear dimensionality reduction techniques such as principal component analysis (PCA) or latent semantic analysis (LSA),

[3] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

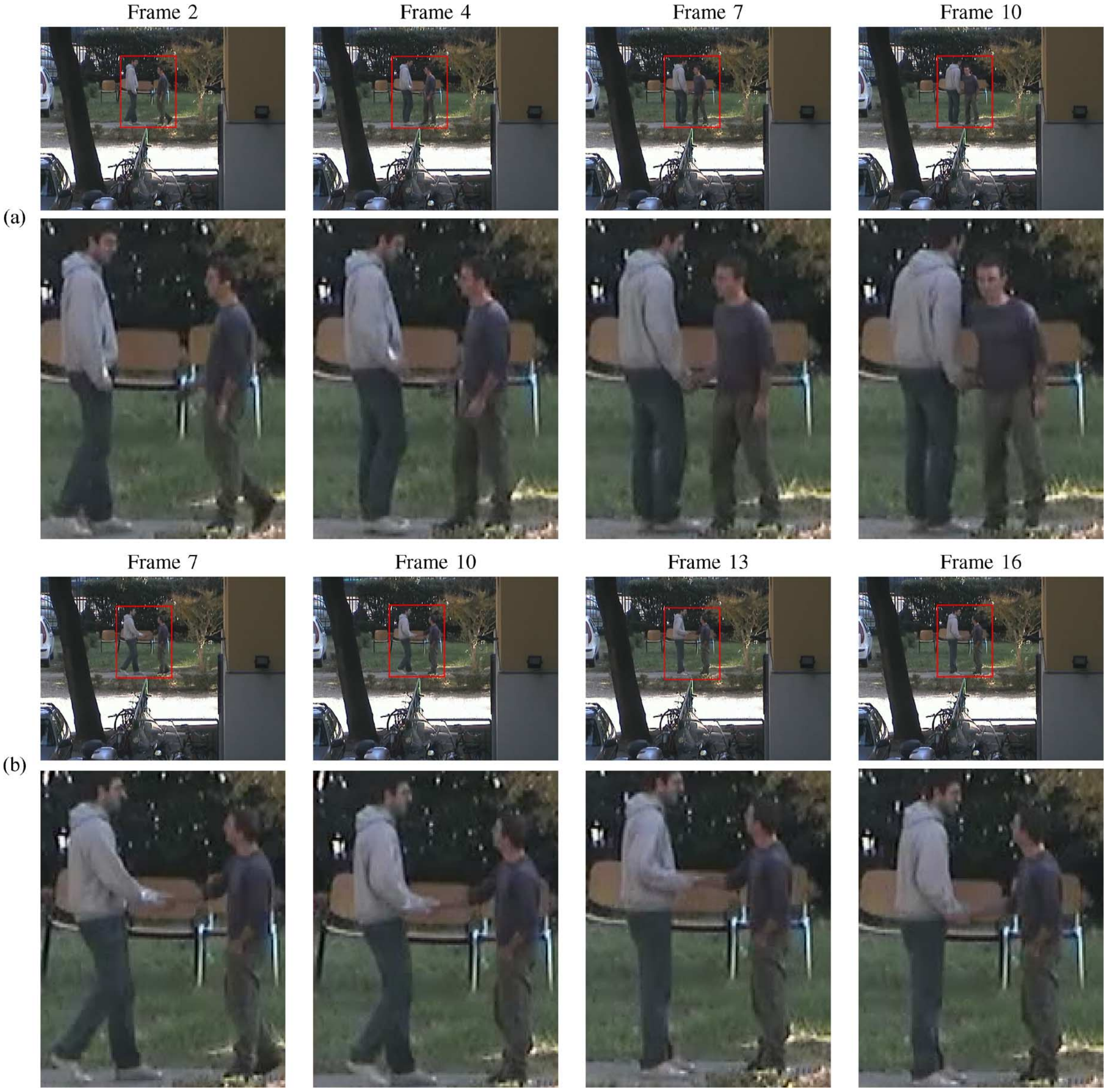


Fig. 10. Sample frames of (a) *give object* and (b) *handshake* action sequences in the MICC-Surveillance dataset. For each sequence, the second row shows the detail indicated in red in the first row.

are not suited to this end because they are not able to handle high-order correlations between codewords that are present in human action representation [53]. We have therefore applied nonlinear dimensionality reduction with deep belief networks (DBNs) [53], [54]. A DBN is composed of several restricted Boltzmann machines (RBM) building blocks that encode levels of nonlinear relationships of the input vectors. It is pre-trained by learning layers incrementally using contrastive divergence [55]. After pre-training, the auto-encoder is built by reversing the network and connecting the top layer of the network to the bottom layer of its reversed version. The auto-encoder is then used to fine-tune the network using a standard back-propagation algorithm.

Since the action representation $H(w)$ can be considered as a coarse probability density estimation of the features of a human action [see (11)], given a set of space-time features $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$, the value of the ith bin of H can be considered as the probability that a space-time descriptor $f \in \mathcal{F}$ is represented by the codeword w_i. This probability can hence be used as an input for an RBM according to [56].

Fig. 11 reports plots of accuracy measured at different codebook sizes, with PCA, LSA, and DBN codebook reduction and radius-based clustering with soft assignment, on the KTH dataset. Codebook reduction was applied to a 4000-codewords codebook. The dimension of the input layer is equal to the size of the uncompressed codebook, and the dimension of the output layer is the compressed codebook size. Each hidden layer is one half the dimension of its input layer. The network depth ranges between five and eight depending on the size of the output codebook. The performance of our approach outperforms that of the method recently proposed in [38], especially for the smaller codebook sizes.

Fig. 12 reports plots of mean computation times for a KTH video sequence as a function of codebook size for radius-based clustering with soft assignment. The accuracy values of Fig. 11 have been reported on the plot for the sake of completeness. It can be noticed that strong codebook size reductions result into time improvements of more than two orders of magnitude. A compressed codebook with 100 codewords scores 89.57% recognition accuracy with respect to 92.66% of a 4000-codewords codebook.

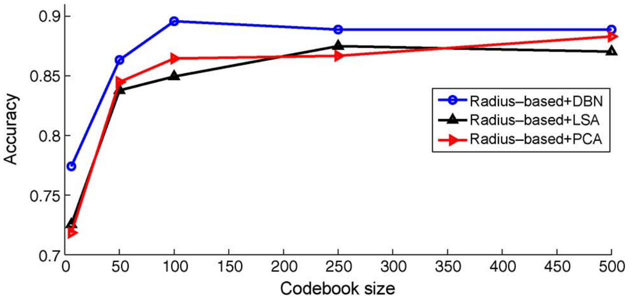


Fig. 11. Classification accuracy on KTH dataset at different codebook sizes, with different codebook reduction techniques, for radius-based clustering with soft assignment.

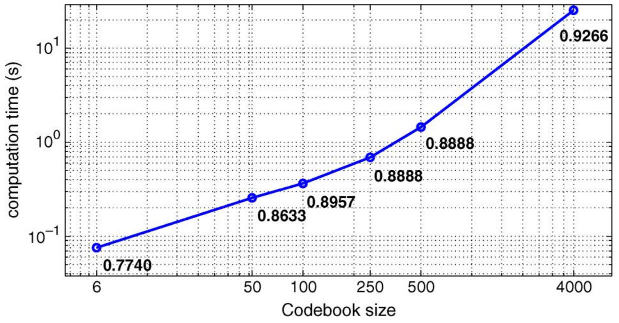


Fig. 12. Mean computation times for a KTH video sequence at different codebook sizes with radius-based clustering and DBNs. The numbers associated to the markers indicate the classification accuracy.

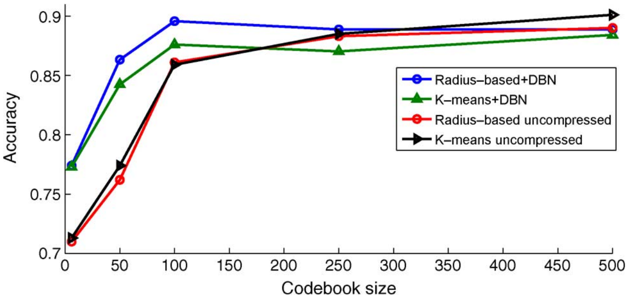


Fig. 13. Classification accuracy as a function of codebook size, for DBN compressed and uncompressed codebooks. Radius-based clustering with soft assignment is compared with k-means clustering with hard assignment.

TABLE VI

Classification Accuracy on MICC-UNIFI Dataset at Different Codebook Sizes, With Different Codebook Reduction Techniques, for Radius-Based Clustering with Soft Assignment. Using PCA and LSA it is not Possible to Create Codebooks Larger Than the Number of Training Videos; Using DBNs This Issue is not Present

Codebook size	6	50	100	250	500
DBN	0.386	0.397	0.412	0.431	0.474
PCA	0.333	0.378	0.405	-	-
LSA	0.330	0.346	0.335	-	-

TABLE VII

Classification of MAP Performance on Hollywood2 Dataset at Different Codebook Sizes, With Different Codebook Reduction Techniques, for Radius-Based Clustering With Soft Assignment

Codebook size	6	50	100	250	500
DBN	0.281	0.372	0.383	0.375	0.374
PCA	0.191	0.323	0.329	0.337	0.338
LSA	0.204	0.322	0.316	0.311	0.314

Fig. 13 shows that, on the one hand, DBN-compressed codebooks provide good accuracy even with very small codebook sizes and, on the other hand, make radius-based clustering still competitive with respect to k-means clustering with 100 or less codewords.

Table VI reports a comparison in terms of classification accuracy at different codebook sizes with DBN, PCA, and LSA on the MICC-UNIFI surveillance dataset. Codebook reduction was applied to the 2000-codeword codebook obtained with radius-based clustering and soft assignment in the previous classification experiment. The smaller number of available training videos, with respect to KTH, is responsible for the reduction in classification accuracy, although the DBNs largely outperform the other methods. This experiment shows another advantage of the use of DBNs over PCA and LSA when the number of sequences available for training is relatively small, i.e., the possibility to create larger dictionaries that usually yield higher classification accuracy although maintaining a speed improvement of an order of magnitude. Table VII reports a comparison of MAP performance obtained using compressed codebooks created with DBN, PCA and LSA on the Hollywood2 dataset. Codebook reduction was applied to the 4000 codeword codebook obtained with radius-based clustering and soft assignment used in the classification experiment. Despite the challenging dataset, the performance is still comparable with that obtained with full-sized codebooks by several approaches reported in [34].

V. Conclusion

In this paper, we have presented a novel method for human action categorization that exploits a new descriptor for spatio-temporal interest points that combines appearance (3-D gradient descriptor) and motion (optic flow descriptor) and effective codebook creation based on radius-based clustering and a soft assignment of feature descriptors to codewords. The approach was validated on KTH and Weizmann datasets, on the Hollywood2 dataset, and on a new surveillance dataset that contain unconstrained video sequences that include more realistic and complex actions. Results outperform the state-of-the-art with no parameter tuning. We have also shown that a strong reduction of computation time can be obtained by applying codebook size reduction with DBNs, with small reduction of classification performance.

References

[1] L. Shao and R. Mattivi, “Feature detector and descriptor evaluation in human action recognition,” in *Proc. CIVR*, 2010, pp. 477–484.

[2] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, “Event detection and recognition for semantic annotation of video,” *Multimedia Tools Applic.*, vol. 51, no. 1, pp. 279–302, 2011.

[3] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” in *Proc. CVPR*, 2003, pp. 264–271.

[4] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, 2007.

[5] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, “The challenge problem for automated detection of 101 semantic concepts in multimedia,” in *Proc. ACM Multimedia*, 2006, pp. 421–430.

[6] A. G. Hauptmann, M. G. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proc. IEEE*, vol. 96, no. 4, pp. 602–622, Apr. 2008.

[7] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understanding*, vol. 104, no. 2–3, pp. 90–126, 2006.

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

[9] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vis. Image Understanding*, vol. 108, no. 1–2, pp. 4–18, 2007.

[10] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[11] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[12] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. ICCV*, 2003, pp. 726–733.

[13] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. CVPR*, 2005, pp. 984–989.

[14] L. Gorelick, M. Blank, E. Schechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[15] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. ICCV*, 2009, pp. 444–451.

[16] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition," in *Proc. CVPR*, 2009, pp. 872–879.

[17] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2–3, pp. 107–123, 2005.

[18] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. VSPETS*, 2005, pp. 65–72.

[19] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. ICPR*, 2004, pp. 32–36.

[20] X. Sun, M. Chen, and A. G. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. CVPR4HB Workshop*, 2009, pp. 58–65.

[21] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. ECCV*, 2008, pp. 650–663.

[22] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT descriptor and its application to action recognition," in *Proc. ACM Multimedia*, 2007, pp. 357–360.

[23] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-Gradients," in *Proc. BMVC*, 2008, pp. 1–10.

[24] Z. Gao, M.-Y. Chen, A. G. Hauptmann, and A. Cai, "Comparing evaluation protocols on the KTH dataset," in *Proc. HBU Workshop*, 2010, pp. 88–100.

[25] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008, pp. 1–8.

[26] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 2169–2178.

[27] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. CVPR*, 2010, pp. 2046–2053.

[28] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Proc. CVPR*, 2009, pp. 1996–2003.

[29] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1/2, pp. 43–72, 2005.

[30] L. Shao, R. Gao, Y. Liu, and H. Zhang, "Transform based spatio-temporal descriptors for human action recognition," *Neurocomputing*, vol. 74, pp. 962–973, 2011.

[31] G. Yu, N. Goussies, J. Yuan, and Z. Liu, "Fast action detection via discriminative random forest voting and top-k subvolume search," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 507–517, Jun. 2011.

[32] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, 2008.

[33] L. Cao, L. Zicheng, and T. Huang, "Cross-dataset action detection," in *Proc. CVPR*, 2010, pp. 1998–2005.

[34] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009, pp. 1–11.

[35] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003, pp. 1470–1477.

[36] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 42–53, Jan. 2010.

[37] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. CVPR*, 2008, pp. 1–8.

[38] Y. Kong, X. Zhang, W. Hu, and Y. Jia, "Adaptive learning codebook for action recognition," *Pattern Recognit. Lett.*, vol. 32, no. 8, pp. 1178–1186, 2011.

[39] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Proc. CVPR*, 2008, pp. 1–8.

[40] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Proc. CVPR*, 2010, pp. 2061–2068.

[41] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. ICCV*, 2005, pp. 604–610.

[42] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. CVPR*, 2009, pp. 2929–2936.

[43] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Effective codebooks for human action categorization," in *Proc. ICCV VOEC*, 2009, pp. 506–513.

[44] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. DARPA IU Workshop*, 1981, pp. 674–679.

[45] K. Mikolajczyk, B. Leibe, and B. Schiele, "Local features for object class recognition," in *Proc. ICCV*, 2005, pp. 1792–1799.

[46] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[47] R. Vezzani and R. Cucchiara, "Video surveillance online repository (ViSOR): An integrated framework," *Multimedia Tools Applic.*, vol. 50, no. 2, pp. 359–380, 2010.

[48] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Proc. CVPR*, 2008, pp. 1–8.

[49] K. Rapantzikos, Y. Avrithis, and S. Kollia, "Dense saliency-based spatiotemporal feature points for action recognition," in *Proc. CVPR*, 2009, pp. 1454–1461.

[50] S.-F. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *Proc. ICCV*, 2007, pp. 1–8.

[51] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. CVPR*, 2009, pp. 1948–1955.

[52] M.-Y. Chen and A. G. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," CMU, 2009.

[53] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction: A comparative review," Tilburg University, Tech. Rep. TiCC-TR 2009-005, 2009.

[54] E. G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[55] M. A. Carreira Perpinan and G. E. Hinton, "On contrastive divergence learning," in *Proc. AISTATS*, 2005, pp. 17–24.

[56] E. G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

Lamberto Ballan (M'11) received the Laurea degree in computer engineering and Ph.D. degree in computer engineering, multimedia and telecommunication from the University of Florence, Florence, Italy, in 2006 and 2011, respectively,

He is a Postdoctoral Researcher with the Media Integration and Communication Center, University of Florence, Florence, Italy. He was a Visiting Scholar with Télécom ParisTech/ENST, Paris, France, in 2010. His main research interests focus on multimedia information retrieval, image and video analysis, pattern recognition, and computer vision. His work was conducted in the context of several EU and national projects, and his results have led to more than 20 publications in international journals and conferences, mainly in multimedia and image analysis.

Dr. Ballan received the Best Paper Award by the ACM-SIGMM Workshop on Social Media in 2010. He coorganized the 1st International Workshop on Web-scale Vision and Social Media in conjunction with ECCV 2012.

Marco Bertini (M'08) received the Laurea degree in electronic engineering and Ph.D. degree from the University of Florence, Florence, Italy, in 1999 and 2004, respectively.

He is currently with the Media Integration and Communication Center, University of Florence, Florence, Italy. His interests are focused on digital libraries, multimedia databases, and social media. On these subjects, he has addressed semantic analysis, content indexing and annotation, semantic retrieval and transcoding. He has authored or coauthored 17 journal papers and more than 75 peer-reviewed conference papers, with h-index: 16 (according to Google Scholar).

Dr. Bertini received the Best Paper Award by the ACM-SIGMM Workshop on Social Media in 2010. He has been involved in five European Union research projects as WP coordinator and researcher. He coorganized the 2010 International Workshop on Multimedia and Semantic Technologies (MUST 2010) and the 2nd IEEE Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS) in conjunction with ICCV 2011. He was chair of the ACM MM 2010 Open Source Software Competition and the 1st International Workshop on Web-scale Vision and Social Media in conjunction with ECCV 2012.

Alberto Del Bimbo (M'90) is a Full Professor of computer engineering with the University of Florence, Florence, Italy, where he is also the director of the Master in Multimedia, and the director of the Media Integration and Communication Center. His research interests include pattern recognition, multimedia information retrieval, computer vision, and human-computer interaction. He has authored or coauthored more than 250 publications in some of the most distinguished scientific journals and international conferences and is the author of the monograph Visual Information Retrieval. He is an associate editor of *Multimedia Tools and Applications, Pattern Analysis and Applications, Journal of Visual Languages and Computing, International Journal of Image and Video Processing,* and *International Journal of Multimedia Information Retrieval.*

Prof. Del Bimbo is an IAPR fellow and was associate editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

Lorenzo Seidenari (S'09) received the Laurea degree in computer engineering and Ph.D. degree in computer engineering, multimedia and telecommunication from the University of Florence, Florence, Italy, in 2008 and 2012, respectively.

Currently, he is a Postdoctoral Researcher with the Media Integration and Communication Center, University of Florence. His main research interests are focused on the application of pattern recognition and machine learning to computer vision and specifically in the field of human activity recognition.

Giuseppe Serra received the Laurea degree in computer engineering and Ph.D. degree in computer engineering, multimedia and telecommunication from the University of Florence, Florence, Italy, in 2006 and 2010, respectively.

He is a Postdoctoral Researcher with the Media Integration and Communication Center, University of Florence, Italy. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, and at Télécom ParisTech/ENST, Paris, France, in 2006 and 2010, respectively. His research interests include image and video analysis, multimedia ontologies, image forensics, and multiple-view geometry. He has authored or coauthored more than 25 publications in scientific journals and international conferences.

Dr. Serra received the Best Paper Award by the ACM-SIGMM Workshop on Social Media in 2010.