Effective Codebooks for Human Action Categorization

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari and Giuseppe Serra
Media Integration and Communication Center, University of Florence, Italy
http://www.micc.unifi.it/vim

Abstract

In this paper we propose a new method for human action categorization by using an effective combination of novel gradient and optic flow descriptors, and creating a more effective codebook modeling the ambiguity of feature assignment in the traditional bag-of-words model. Recent approaches have represented video sequences using a bag of spatio-temporal visual words, following the successful results achieved in object and scene classification. Codebooks are usually obtained by k-means clustering and hard assignment of visual features to the best representing codeword. Our main contribution is two-fold. First, we define a new 3D gradient descriptor that combined with optic flow outperforms the state-of-the-art, without requiring fine parameter tuning. Second, we show that for spatio-temporal features the popular k-means algorithm is insufficient because cluster centers are attracted by the denser regions of the sample distribution, providing a non-uniform description of the feature space and thus failing to code other informative regions. Therefore, we apply a radius-based clustering method and a soft assignment that considers the information of two or more relevant candidates. This approach generates a more effective codebook resulting in a further improvement of classification performances. We extensively test our approach on standard KTH and Weizmann action datasets showing its validity and outperforming other recent approaches.

1. Introduction and previous work

Automatic human activity recognition methods are useful for many applications such as video surveillance, video annotation and retrieval and human-computer interaction. For example, in video surveillance, an automatic action classification system that alerts an operator of a possible dangerous situation can reduce human effort and mistakes.

However, building a general human activity recognition and classification system is a challenging problem, because of the variations in environment, people and actions. In fact environment variation can be caused by cluttered or moving background, camera motion, illumination changes. People may have different size, shape and posture appearance. Semantically equivalent actions can manifest differently or partially; for example, imagine the different ways of running or actions that can be only partially observed due to occlusions.

Over the past decade, this problem has received considerable attention. Existing action recognition approaches can be classified as using *holistic information* or *part-based information*. An early work based on holistic representation was proposed by Bobick *et al*. [1]. They proposed the motion history images, 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. This representation can be matched using global statistics, such as moment features. Although this method is efficient, it is assumed to have a well segmented foreground and background. Efros *et al*. [5] created stabilized spatio-temporal volumes for each object whose action is to be classified. For each volume a smoothed dense optic flow field is extracted and used as descriptor. This method is particularly suited for distant objects where detailed information of the appearance is unavailable. Yilmaz and Shah [29] used a spatio-temporal volume, built stacking object regions obtained by a contour tracking method, in consecutive frames. Descriptors encoding direction, speed and local shape of the resulting surface are generated by measuring local differential geometrical properties. Gorelick *et al*. [9] analysed three-dimensional shapes induced by the silhouettes and exploited the solution to the Poisson equation to extract features, such as shape structure and orientation. These methods require robust tracking to generate the 3D volumes. Moreover most of the holistic-based approaches are computationally expensive due to the requirement of pre-processing the input data (e.g. to perform background subtraction, shape extraction, optic flow calculation, object tracking) and they perform better in a controlled environment.

Part-based representations, that exploit interest point detectors combined with robust descriptor methods, have been

used very successfully for object and scene classification tasks [7, 23, 28, 30]. Recently, part-based models have been successfully applied to the human action classification problem, because they overcome some limitations of holistic models such as the necessity of performing background subtraction and tracking. Laptev [12] proposed an extension to the Harris-Förstner corner detector for the spatio-temporal case; interesting parts are extracted from voxels surrounding local maxima of spatio-temporal corners, i.e. locations of videos which exhibit strong variations of intensity both in spatial and temporal directions. The extension of the scale-space in the temporal dimension yields a method for automatic scale-selection. Schüldt *et al*. [21] successfully used these features for human action classification by discretizing them into codewords and producing an histogram of the occurring words for each shot. Dollár *et al*. [4] have followed in principle the same approach of Laptev, but suggested to treat time differently from space and to look for locally periodic motion using a quadrature pair of Gabor filters. Their approach produces a denser sampling of the spatio-temporal volume but does not provide a scale-selection criterion. Comparison of the experimental results w.r.t. the approach of Schüldt *et al*. shows an improvement on the same dataset. Niebles *et al*. [18] have then trained an unsupervised probabilistic topic model on the same features as Dollár *et al*., obtaining comparable classification performance. More recently, Laptev *et al*. [13] have addressed the human action recognition problem in more realistic video settings. They also abandon the scale selection approach, preferring a structural representation based on dense temporal and spatial scale sampling inspired by spatial pyramids [14], showing an improvement of the state-of-the-art results. Finally, Willems *et al*. [26] proposed a new efficient and scale-invariant spatio-temporal detector and descriptor, extending the static SURF features.

All of these part-based approaches use the codebook paradigm that allows classification by describing a video as a bag of words, where video features are represented by discrete visual codewords. These are defined beforehand in a given vocabulary. A vocabulary, in the object and scene classification domain, is commonly obtained by following one of two approaches: an *annotation approach* [25] or a *data-driven approach* [2, 23, 30]. The annotation approach obtains a vocabulary by assigning meaningful labels to image patches (e.g. sky, water, vegetation, etc.) while, in contrast, a data-driven approach applies vector quantization on the features using typically k-means clustering. However, despite of its popularity, this is not the optimal solution. Jurie and Triggs [10] have shown that in k-means clustering the centres are almost exclusively around the denser regions in descriptor space and thus fail to code other informative regions. They show that k-means works well for texture analysis in homogeneous images, but the images that arise in natural scenes have far less uniform statistics. For this reason they proposed a scalable acceptance radius-based clustering that generates better codebooks. Nevertheless, all the previous part-based methods for human action recognition use the k-means algorithm for codebook creation. To the best of our knowledge, few papers address approaches to obtain an efficient codebook in human action recognition area. Liu *et al*. [16] proposed a method to automatically find the optimal number of word clusters by utilizing maximization of mutual information (MMI) between words and actions. Initially they apply k-means and then MMI clustering is used to discover a compact representation from the initial codebook of words. They show an improvement of the performance with the learned optimal number of words. A different approach has been proposed by Mikolajczyk and Uemura [17] that recently explored the idea of using a large number of features represented in many vocabulary trees instead of a single flat vocabulary.

Independently of the clustering algorithm, one of the main drawback of the codebook approach, recently pointed out in object and scene classification, is the hard assignment of image feature vectors to codewords in the vocabulary [19, 24]. This hard assignment is particularly critical because of two main issues. The first one (*uncertainty*) refers to the problem of selecting the correct codeword out of two or more relevant candidates; the second one (*plausibility*) denotes the problem of selecting a codeword without a suitable candidate in the vocabulary.

In this paper we describe a new method for classification of human actions that relies on an appropriate quantization method, dealing with the ambiguity of the traditional codebook model. Our main contribution is two-fold: *i)* the definition of gradient and optic flow descriptors that, combined together, outperform the state-of-the-art without requiring fine parameter tuning; *ii)* a radius-based clustering method and a soft assignment procedure that, considering the information of two or more relevant candidates, are able to generate effective codebooks showing a further improvement of classification performances. The rest of the paper is organized as follows. The interest point detector and descriptors are presented in the next section. The techniques for action representation and categorization, including the codebook creation, are discussed in Sect. 3. Experimental results, with an extensive comparison with state-of-the-art approaches, are discussed in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2. Detector and descriptors

Following the approach commonly used for local interest points in images, the detection and description of spatio-temporal interest points are separated in two different steps. Among the different spatio-temporal interest point detectors available, the spatio-temporal corner detector proposed by

Laptev *et al*. [12] provides a too sparse representation of the actions. For this reason the spatio-temporal interest points detector proposed by Dollár *et al*. [4], that is able to detect a greater number of points, has received a large attention from the scientific community and has been adopted in several recent works [16, 18].

Detector. In our approach we have adopted the detector proposed by Dollár *et al*. [4]. This detector applies two separate linear filters to spatial and temporal dimensions, respectively. The response function is computed as follows:

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

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 $\omega = 4/\tau$, and they give a strong response to the temporal intensity changes, in particular for periodic motion patterns. The interest points are detected at locations where the response is locally maximum.

The main problem of this detector is the fact that it does not cope with scale selection. However, both spatial and temporal scales have to be considered when analyzing motion activity. The spatial scale is related to the ability to detect more or less detailed visual features, while the temporal scale is related to the ability to detect actions that are performed at different speed. In order to cope with the lack of scale selection we run the detector over a set of spatial and temporal scales, to permit the recognition of the same action at different distance and velocity. In particular the spatial scales used are $\sigma = \{2, 4\}$ and the temporal scales are $\tau = \{2, 4\}$. This approach has also some other desirable properties such as a reduced computational complexity w.r.t. scale selection and the production of a richer description of the scene, using a larger number of interest points.

Descriptors. For each detected point a patch that contains the volume that contributed to the response function is considered. The volume is proportional to the scale at which the interest point is detected. Each volume is divided in equally sized sub-regions, three for the spatial dimensions and two for the temporal dimension. To obtain a representation for each spatio-temporal volume, we evaluate a descriptor based on gradients on x, y and t direction and an optic flow descriptor, considering also their combinations. This is motivated by the fact that these two descriptors encode different information. In fact the descriptor based on gradient encodes mostly the visual appearance of each volume, while the optic flow descriptor encodes the motion information. The two descriptors are presented in the following.

The gradient magnitude and orientations in 3D are:

$$M_{3D} = \sqrt{G_x^2 + G_y^2 + G_t^2}, \quad (2)$$

$$\phi = \tan^{-1}(G_t/\sqrt{G_x^2 + G_y^2}), \quad (3)$$

$$\theta = \tan^{-1}(G_y/G_x). \quad (4)$$

where G_x, G_y and G_z are respectively computed using finite difference approximations: $I(x+1,y,t)-I(x-1,y,t)$, $I(x,y+1,t)-I(x,y-1,t)$ and $I(x,y,t+1)-I(x,y,t-1)$. We compute two separated orientation histograms quantizing ϕ and θ and weighting them by the magnitude M_{3D}. The ϕ (with range, $-\frac{\pi}{2}, \frac{\pi}{2}$) and θ ($-\pi$,π) are quantized in four and eight bins respectively. The spatio-temporal gradient is computed after smoothing the values extracted with those of two adjacent scales, to increase the robustness of the feature description. The overall dimension of the descriptor is thus $3 \times 3 \times 2 \times (8+4) \times 2 = 432$. This construction of the three-dimensional histogram is inspired by the approach proposed by Scovanner *et al*. [22], in which they construct a weighted three-dimensional histogram normalized by the solid angle value (instead of quantizing separately the two orientations) to avoid distortions due to the polar coordinate representation. Moreover we do not reorient the 3D neighbourhood, since rotational invariance, which is invaluable in object detection and recognition, is not desired in an action categorization context. We have found that our method is computationally less expensive, equally effective in describing motion information given by appearance variation, and showing a better performance (see comparison results in Tab. 2).

The optic flow is estimated using the Lucas&Kanade algorithm. Considering the optic flow computed for each couple of consecutive frames, the relative apparent velocity of each pixel is (V_x, V_y). These values are expressed in polar coordinates as in the following:

$$M_{2D} = \sqrt{V_x^2 + V_y^2}, \quad (5)$$

$$\theta = \tan^{-1}(V_y/V_x). \quad (6)$$

We compute position dependent histograms as in the gradient based descriptor but, being the optic flow two dimensional, only a single orientation histogram is stored for each of the 18 sub-regions within the voxel. Every sample is weighted with the magnitude M_{2D}, as is done for the gradient-based descriptor. Then we have also added an extra "no-motion" bin that, in our initial experiments, has shown to greatly improve the performance. Thus the final descriptor size is $3 \times 3 \times 2 \times (8+1) = 162$.

We have finally analysed two possible combinations of these descriptors: *i)* a weighted concatenation of the two

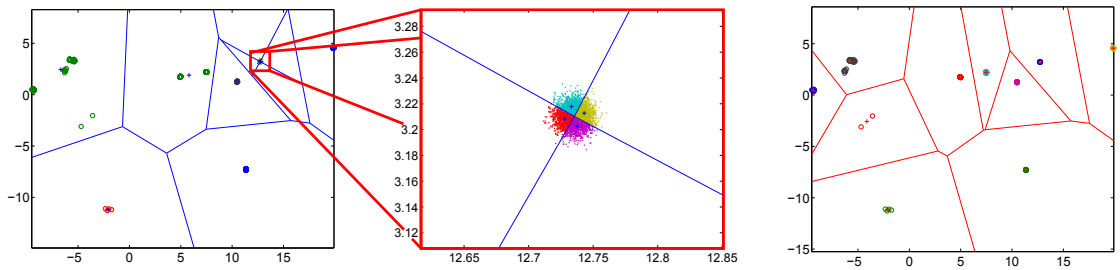


Figure 1. Comparison of k-means and radius-based clustering on a synthetic dataset. (a) k-means clustering; (b) k-means clustering: detail of a dense region that has been split in four clusters; (c) Radius-based clustering.

descriptors and *ii)* a concatenation of the histograms of the bag-of-words that have been computed from the 3D gradient descriptor and from the histogram of optic flow. In the first case the visual words, created according to the bag-of-words paradigm, are computed from a vector that has higher dimensionality, while in the second case the visual words are computed differently for each descriptor and the SVM classifiers are able to pick the best combinations of features, practically resulting in an implicit feature selection.

3. Action Representation and Categorization

The spatio-temporal bag-of-words (BoW) model is built through the creation of a discrete visual vocabulary (or codebook) and then by assigning each feature to the corresponding codeword. First of all, it is required to perform a vector quantization for large sets of feature vectors in a high dimensional space. Typically this is performed through clustering methods and the most common approach is the use of k-means clustering, because of its simplicity and convergence speed. The BoW approach then assigns each feature to the closest vocabulary word and a histogram of visual word frequencies is computed. The histogram is fed to a classifier to predict the action category. The performance of this model depends on the quantization method and on the number of words that are selected.

3.1. Codebook Formation

The use of k-means clustering has some disadvantages: *i)* the cluster centers are attracted by the denser regions of the sample distribution, resulting more clustered near these regions and more sparse otherwise, thus providing a more imprecise quantization for the vectors laying in these latter regions [10]. This effect, due to the assumption of uniform distribution of the features in the descriptor space, is even more pronounced in high dimensional spaces such as those spanned by the spatio-temporal descriptors; a representation of this effect can be obtained visualizing a Voronoi tessellation of the feature space, where Voronoi cells do not uniformly cover the feature space as shown in Fig. 1. Other disadvantages are: *ii)* the clustering is not very robust w.r.t. outliers, *iii)* the number of visual words has to be known in advance, requiring an empirical evaluation of this number.

Radius-based clustering. In order to overcome the limitations of k-means clustering, we explore the idea of using an on-line radius-based clustering technique following a mean-shift approach [3, 8]. In fact, as shown by Jurie and Triggs [10], in the case of dense sampling image representations, it is better to apply a radius-based clustering method. This observation is interesting also for the human action domain because, as previously introduced (Sect. 2), the spatio-temporal features extracted by the Dollár detector [4] can be considered as a dense representation; this fact is even more pronounced using our multi-scale approach. In this case the non-uniformity in the descriptor space, caused by densely sampled patches, is better coded using a radius-based method that is able to allocate centers more uniformly. An example of this effect is shown in Fig. 1 c.

The algorithm starts with an uniform random subsampling D_n of the original dataset points D. Given a radius R, mean shift clustering on D_n is used to find the modes of the samples distribution. A new cluster center is then allocated on the mode corresponding to the maximal density region. Data points of the original dataset D, within a distance less than R from the center, are considered members of this cluster and eliminated for the following iterations. This elimination prevents the algorithm from repeatedly assigning centers to the same high density regions. Finally, the algorithm can be stopped when a "sufficiently" large number of clusters (words) has been identified.

Visual words statistics. One of the assumption in text categorization methods is that, given a natural language textual corpus, the words frequency distribution follows the well-known *Zipf's law*. This is a critical point because, considering this empirical evidence, we can consider words at intermediate frequencies as the most informative for classification. Therefore it is interesting to see how the visual

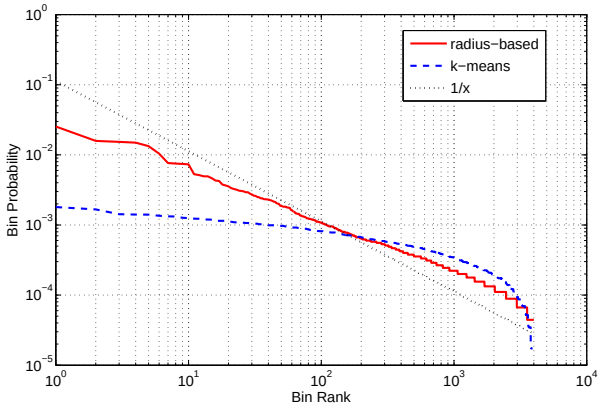


Figure 2. Log-log plots of visual words frequency using k-means and radius-based quantization.

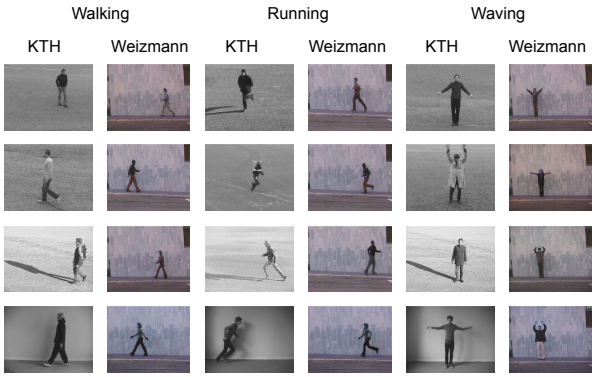


Figure 3. Sample frames from the KTH and Weizmann datasets (Walking, Running and Waving actions).

words are distributed in a visual corpus, as also noted in [10, 20, 28]. In particular, we want to know whether their distribution satisfies Zipf's law. Fig. 2 shows the statistics of visual words frequency using k-means and radius-based quantization on our experimental dataset (see Sect. 4 for details). An "ideal" Zipf's distribution must be a straight line in log-log scale. The figure shows that the distribution of visual words obtained by k-means quantization satisfies the Zipf's law only roughly. In fact, most of the bins has similar frequencies and they are distributed more evenly with respect to the expected power law. In contrast, the proposed radius-based quantization shows a statistics that fit better the expected distribution. This confirms the assumptions discussed in the previous paragraph and confirms that this approach models better medium density frequencies.

3.2. Codeword Assignment

Given a vocabulary, the traditional codebook approach represents a video sequence containing an action by a histogram of codeword frequencies. In particular, for each word w in the vocabulary V the frequency distribution in a sequence is computed by:

$$FD(w) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1 & \text{if } w = \underset{v\in V}{\operatorname{argmin}}(D(v,p_i));\\ 0 & \text{otherwise;}\end{cases} \tag{7}$$

where n is the number of spatio-temporal patches in a sequence, p_i is the i^{th} spatio-temporal patch, and $D(v,p_i)$ is the distance (usually Euclidean) between the codeword v and the patch p_i. This hard assignment, that takes account only of the closest codeword, lacks to consider two issues: codeword *uncertainty* (selection of the correct codeword when two or more candidates are relevant) and codeword *plausibility* (selection of a codeword when all codewords are too far and not representative). We observe that, in our case, the plausibility is less problematic, because the radius-based clustering method that we employ is able to allocate the centers more uniformly. On the other hand, as noted by van Gemert *et al.* [24], in a high-dimensional feature space the codeword uncertainty issue becomes very urgent. In fact, if we consider a codeword as a high-dimensional sphere in feature space, most feature points in this sphere lay near the surface and are close to the boundary between different codewords. For this reason the distribution of the codewords in a sequence has to contain the information of two or more relevant candidates. This can be done by smoothing the hard assignment of a spatio-temporal patch to the codeword vocabulary using Gaussian kernel density estimation, computing the uncertainty frequency distribution with:

$$UFD(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{K_\sigma(D(w,p_i))}{\sum_{j=1}^{|V|}K_\sigma(D(v_j,p_i))} \tag{8}$$

where D is the Euclidean distance and K_σ is the Gaussian kernel:

$$K_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) \tag{9}$$

where σ is the scale parameter of the Gaussian kernel; this parameter has to be tuned on the training set, because dependent on the dataset, the features length and their range values.

4. Experimental Results

We tested our approach on two datasets commonly used for human action recognition: the KTH and Weizmann datasets. The KTH dataset contains 2391 video sequences showing six actions: walking, running, jogging, hand-clapping, hand-waving, boxing. They are performed by 25 actors under four different scenarios of illumination, appearance and scale change. The video resolution is 160×120 pixel. The Weizmann dataset contains 93 video sequences showing nine different people, each performing

ten actions such as 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. The video resolution is 180×144 pixel. An example of the differences between the two dataset is shown in Fig. 3, where sample frames selected from videos containing the same action in the two sets are compared each other.

Two approaches were followed during the training phase, due to the different sizes in the datasets. The SVM classifiers used for the KTH dataset were trained on videos of 16 actors and the performance was evaluated using the videos of the remaining 9 actors. Measures have been taken according to a five-fold cross-validation. Due to the small size of the Weizmann dataset the classifiers were trained on actions from eight actors and tested on the remaining one. Measures have been taken using the leave-one-out cross-validation. This setup is identical to the most recent works in action recognition domain and thus is suitable for a direct comparison [11, 13, 18, 27]. Classification is performed using non-linear SVMs with the χ^2 kernel [30]:

$$K_{\chi^2}(p,q) = \exp\left(-\frac{1}{2\gamma}\sum_{k=1}^{N}\frac{(p_k - q_k)^2}{(p_k + q_k)}\right) \qquad (10)$$

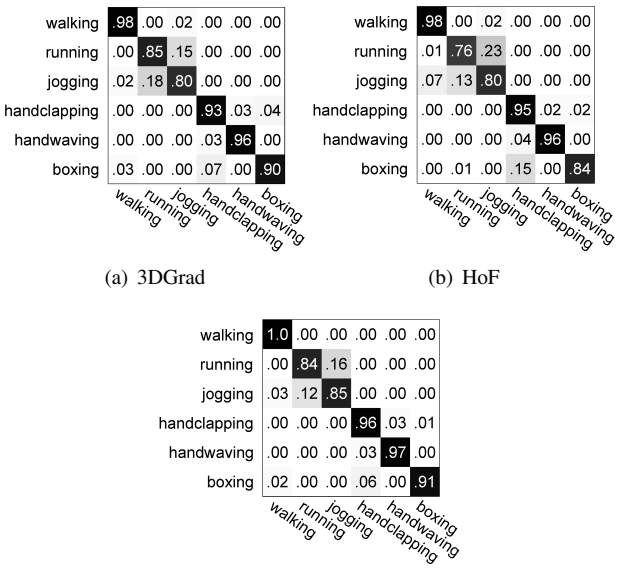
where N is the vocabulary size, p and q are histograms of word occurrences. The value of the kernel parameter γ is obtained by cross-validation on the training set. For multi-class classification, we use the *one-vs-one* approach.

4.1. Evaluation of our descriptor

Table 1 evaluates the performance of our proposed descriptors, comparing the performance of each descriptor alone and the two possible combinations discussed in Sect. 2. The experiments have been carried on using the setup described above, and the quantization approach used is k-means clustering (using 4000 words for KTH and 1000 for Weizmann), in order to be directly comparable with other approaches. In the first two rows we report results obtained using only one of the two descriptors, 3D gradient in the first row and histogram of optic flow in the second. In the third row are reported the results for the descriptor that is obtained through a weighted concatenation of the two descriptors, while in row four the descriptor is composed by the concatenation of the histograms of the bag-of-words that have been computed from the 3D gradient descriptor and from the histogram of optic flow. The best result, on both datasets, is achieved by the concatenation of the histograms of the BoWs computed from both descriptors. This is due to the fact that the performance of 3D gradient and HoF are quite complementary (see Fig. 4). For example, the action recognition performance for the boxing class on the KTH dataset is better when using the 3D gradient instead of the HoF description, while for handclapping is the opposite

Descriptor	KTH	Weizmann
3DGrad	90.38 ± 0.8	92.30±1.6
HoF	88.04 ± 0.7	89.74±1.8
3DGrad_HoF combination	91.09 ± 0.4	92.38±1.9
3DGrad+HoF combination	**92.10 ± 0.4**	**92.41 ±1.9**

Table 1. Comparison of our descriptors, alone and combined, on the KTH and Weizmann datasets.



(a) 3DGrad (b) HoF

(c) 3DGrad+HoF combination

Figure 4. Confusion matrices on the test set KTH actions.

case. It can be observed (Fig. 4 c) that the concatenation of the histograms improves the performance for all the classes except one, running class. In the Weizmann dataset we obtain a smaller improvement, with the concatenation of histogram, probably caused by the smaller training set that is available and the increased size of the representation.

4.2. Performances obtained by effective codebooks

In this set of experiments we evaluate the different codebook creation approaches presented in Sect. 3. The datasets used are the KTH and Weizmann with the same experimental setup described above, and the descriptor is the concatenation of the histograms of bag-of-words computed from 3D gradient and optic flow descriptors (3DGrad+HoF). Fig. 5 compares the classification performances obtained by the standard k-means and hard assignment approach - commonly used by previous works - with the proposed radius-based clustering and soft assignment. The graph reports the variation in accuracy w.r.t. the number of visual words, up to the number of words (4000 for KTH, 1000 for Weizmann) that were used in the previous experiments.

With a very low number of words the soft radius-based clustering method has a lower performance than k-means, since in this approach the words that are used are those that

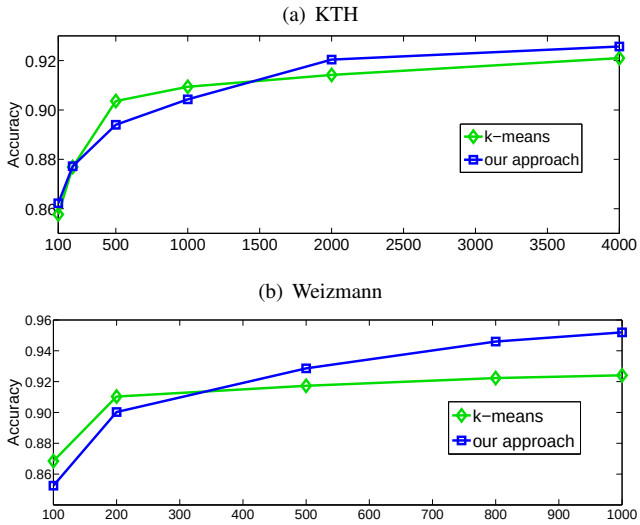


Figure 5. Comparison of classification accuracies on KTH and Weizmann datasets using the combined descriptor (3DGrad+HoF) and *i)* k-means based codebooks and *ii)* our effective codebooks approach (i.e. radius-based clustering + soft assignment).

are more common (i.e. those that provide less discriminative information). However, this effect disappears rapidly (above 1500 words for KTH and 400 words for Weizmann) due to the more effective choice of the words, as discussed in Sect. 3.2. The radius-based clustering extended so as to account for codeword uncertainty outperforms k-means clustering and classification results are improved in both datasets; in particular, it has a better performance even with a relatively low number of visual words (e.g. $\sim$ 2000 for KTH and $\sim$ 500 for Weizmann). Indeed the radius-based clustering better encodes sparser regions while the soft assignment is able to moderates uncertainty in the denser ones, leading thus to more effective codebooks.

We report on Fig. 6 the final classification performance on KTH and Weizmann datasets, obtained using the proposed soft radius-based quantization, as confusion matrices. Interestingly, the major confusion occurs between similar classes (running-jogging on KTH and jump-skip on Weizmann). The overall accuracy on KTH is 92.57% while on Weizmann is 95.41%.

4.3. Comparison to state-of-the-art

In Table 2 we report a comparison of the average class accuracy of our approach with state-of-the-art results, reported by other researchers.

Results obtained on KTH using our combined descriptor (3DGrad+HoF) united with the proposed effective codebook formation outperform previous works based on standard BoW models [21, 4, 11, 13, 18, 26, 27], even those that employ fine tuning of parameters or additional structural descriptors. Note that the previous state-of-the-art result (91.8%), achieved by Laptev *et al*. [13] using their best

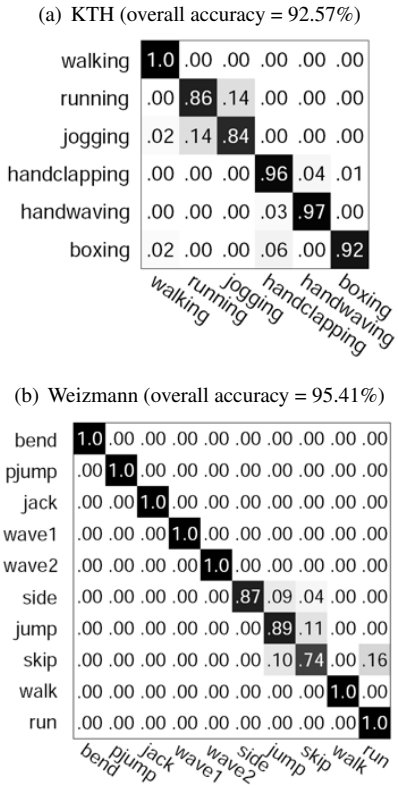


Figure 6. Final confusion matrices on KTH and Weizmann.

combination of features, is obtained performing a greedy search on different combination of descriptors (HoG and HoF) and grids, which add structural information. Our results outperform also those of Kläser *et al*. [26] (91.4%) that use a single 3D gradient descriptor but with a heavy optimization of eight descriptor parameters, resulting in a high dependence on the dataset used.

Also when considering the Weizmann dataset we outperform previous BoW-based works [11, 18, 22], and also the results reported by Liu *et al*. [15] (90.4%) obtained combining and weighting multiple features. However, we cannot compare to results by Gorelick *et al*. [9] or Fathi and Mori [6] because they use an holistic representation and more data given by segmentation masks.

5. Conclusions

In this paper we have presented a novel method for human action categorization based on a new descriptor for spatio-temporal interest points, that combines appearance (3D gradient descriptor) and motion (optic flow descriptor), and on an effective codebook formation. We replaced the traditional codebook quantization method using a radius-based clustering algorithm and a soft assignment of features descriptors to codewords. The approach was validated on two popular datasets (KTH and Weizmann), showing results that outperform state-of-the-art BoW approaches, without

Method	KTH	Weizmann
Our method	**92.57**	**95.41**
Laptev *et al.* [13]	91.8	-
Dollár *et al.* [4]	81.2	-
Wong and Cipolla [27]	86.62	-
Scovanner *et al.* [22]	-	82.6
Niebles *et al.* [18]	83.33	90
Liu *et al.* [15]	-	90.4
Kläser *et al.* [11]	91.4	84.3
Willems *et al.* [26]	84.26	-
Schüldt *et al.* [21]	71.7	-

Table 2. Comparison of our method to state-of-the-art.

requiring parameter tuning employed by the previous best results. The proposed approach is modular and each contribution of this paper can be adapted to any framework based on interest points and BoW. Our future work will deal with evaluation on real world videos.

Acknowledgements. This work is partially supported by the EU IST VidiVideo Project (Contract FP6-045547) and IM3I Project (Contract FP7-222267).

References

[1] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 1

[2] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008. 2

[3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 4

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. of VSPETS*, 2005. 2, 3, 4, 7

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. of ICCV*, 2003. 1

[6] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *Proc. of CVPR*, 2008. 7

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of CVPR*, 2003. 2

[8] B. Georgescu, I. Shimshoni, and P. Meer. Mean shift based clustering in high dimensions: A texture classification example. In *Proc. of ICCV*, 2001. 4

[9] L. Gorelick, M. Blank, E. Schechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12), 2007. 1, 7

[10] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. of ICCV*, 2005. 2, 4, 5

[11] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. *Proc. of BMVC*, 2008. 6, 7

[12] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005. 2, 3

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. of CVPR*, 2008. 2, 6, 7

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of CVPR*, 2006. 2

[15] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *Proc. of CVPR*, 2008. 7

[16] J. Liu and M. Shah. Learning human actions via information maximization. In *Proc. of CVPR*, 2008. 2, 3

[17] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. In *Proc. of CVPR*, 2008. 2

[18] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008. 2, 3, 6, 7

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. of CVPR*, 2008. 2

[20] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007. 5

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc. of ICPR*, 2004. 2, 7

[22] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. of ACM Multimedia*, 2007. 3, 7

[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*, 2003. 2

[24] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *Proc. of ECCV*, 2008. 2, 5

[25] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007. 2

[26] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. of ECCV*, 2008. 2, 7

[27] S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *Proc. of ICCV*, 2007. 6, 7

[28] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proc. of MIR*, 2007. 2, 5

[29] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *Proc. of CVPR*, 2005. 1

[30] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007. 2, 6